Dynamics of transcription factor binding site evolution

Murat Tuğrul*, Tiago Paixão, Nicholas H. Barton, Gašper Tkačik
Institute of Science and Technology Austria, Am Campus 1, A-3400 Klosterneuburg, Austria

(Dated: November 9, 2015)

Evolution of gene regulation is crucial for our understanding of the phenotypic differences between species, populations and individuals. Sequence-specific binding of transcription factors to the regulatory regions on the DNA is a key regulatory mechanism that determines gene expression and hence heritable phenotypic variation. We use a biophysical model for directional selection on gene expression to estimate the rates of gain and loss of transcription factor binding sites (TFBS) in finite populations under both point and insertion/deletion mutations. Our results show that these rates are typically slow for a single TFBS in an isolated DNA region, unless the selection is extremely strong. These rates decrease drastically with increasing TFBS length or increasingly specific protein-DNA interactions, making the evolution of sites longer than ~ 10 bp unlikely on typical eukaryotic speciation timescales. Similarly, evolution converges to the stationary distribution of binding sequences very slowly, making the equilibrium assumption questionable. The availability of longer regulatory sequences in which multiple binding sites can evolve simultaneously, the presence of "pre-sites" or partially decayed old sites in the initial sequence, and biophysical cooperativity between transcription factors, can all facilitate gain of TFBS and reconcile theoretical calculations with timescales inferred from comparative genomics.

Author Summary

Evolution has produced a remarkable diversity of living forms that manifests in qualitative differences as well as quantitative traits. An essential factor that underlies this variability is transcription factor binding sites, short pieces of DNA that control gene expression levels. Nevertheless, we lack a thorough theoretical understanding of the evolutionary times required for the appearance and disappearance of these sites. By combining a biophysically realistic model for how cells read out information in transcription factor binding sites with model for DNA sequence evolution, we explore these timescales and ask what factors crucially affect them. We find that the emergence of binding sites from a random sequence is generically slow under point and insertion/deletion mutational mechanisms. Strong selection, sufficient genomic sequence in which the sites can evolve, the existence of partially decayed old binding sites in the sequence, as well as certain biophysical mechanisms such as cooperativity, can accelerate the binding site gain times and make them consistent with the timescales suggested by comparative analyses of genomic data.

INTRODUCTION

Evolution produces heritable phenotypic variation within and between populations and species on relatively short timescales. Part of this variation is due to differences in gene regulation, which determines how much of each gene product exists in every cell. These gene expression levels are heritable quantitative traits subject to natural selection [1–3]. While the importance of their variability for the observed phenotypic variation is still debated [4], it is believed to be crucial within closely related species or in populations whose proteins are functionally or structurally similar [5]. The genetic basis for gene expression differences is thought to be non-coding regulatory DNA, but our understanding of its evolution is still immature; this is due, in part, to the lack of precise knowledge about the mapping between the regulatory sequence and the resulting expression levels.

Transcriptional regulation is the most extensively studied mechanism of gene regulation. Transcription factor proteins (TFs) recognize and bind specific DNA sequences called binding sites, thereby affecting the expression of target genes. Eukaryotic regulatory sequences, i.e., enhancers and promoters, are typically between a hundred and several thousand base pairs (bp) in length [6], and can harbor many transcription factor binding sites (TFBSs), each typically consisting of $6-12$ bp. The situation is different in prokaryotes: they lack enhancer regions and have one or a few TFBSs which are typically longer, between 10 to 20 bp in length [7, 8]. Differences in TF binding are thought to arise primarily due to changes in the regulatory sequence at the TF binding sites rather than changes in the cellular

* mtugrul@ist.ac.at

environment or the TF proteins themselves [10]. Nevertheless, a theoretical understanding of the relationship between the evolution of the regulatory sequence and the evolution of gene expression levels remains elusive, mostly because of the complex interaction of evolutionary forces and biophysical processes [11].

From the evolutionary perspective, the crucial question is whether and when these regulatory sequences can evolve rapidly enough so that new phenotypic variants can arise and fix in the population over typical speciation timescales. Comparative genomic studies in eukaryotes provide evidence for the evolutionary dynamics of TF binding, highlighting the possibility for rapid and flexible TFBS gain and loss between closely related species on timescales of as little as a few million years [12, 13]. Examples include quick gain and loss events that cause divergent gene expression [14], or the compensation of such events by turn-over at other genome locations [15]; gain and loss events sometimes occur even in the presence of strong constraints on expression levels [16, 17]. Furthermore, such events enabled new binding sites on sex chromosomes that arose as recently as $1-2$ million years ago [18, 19]. There are examples of rapid regulatory DNA evolution across and within populations requiring shorter timescales, i.e. $10.000-100.000$ years [2, 20–22]. On the other hand, strict conservation has also been observed at orthologous regulatory locations even in distant species (e.g., [23]). Taken together, these facts suggest that the rates of TFBS evolution can extend over many orders of magnitude and differ greatly from the point mutation rate at a neutral site. To study the evolutionary dynamics of regulatory sequences and understand the relevant timescales, we set up a theoretical framework with a special focus on the interplay of both population genetic and biophysical factors, briefly outlined below.

Sequence innovations originate from diverse mutational mechanisms in the genome. While tandem repeats [24] or transposable elements [25] may be important in evolution, the better studied and more widespread mutation types still need to be better understood in the context of TFBS evolution. Specifically, we ask how the evolutionary dynamics are affected by single nucleotide (point) mutations, as well as by insertions and deletions (indels). New mutations in the population are selected or eliminated by the combined effects of selection and random genetic drift. Although the importance of selection [26–28] and mutational closeness of the initial sequences [29, 30] for TF binding site evolution has already been reported, the belief in fast evolution via point mutations without selection (i.e., neutral evolution) persists in the literature (e.g.,[5, 13]), mainly due to Stone & Wray's (2001) misinterpretation of their own simulation results [31] (see Macarthur & Brookfield (2004) [29]). This likely reflects the current lack of theoretical understanding of TFBS evolution in the literature, even under the simplest case of directional selection. Basic population genetics shows that directional selection is expected to cause a change, e.g., yield a functional binding site, over times on the order of $1/(NsU_b)$, where N is the population size, s is the selection advantage of a binding site, and U_b is the beneficial mutation rate [32]. This process can be extremely slow, especially under neutrality, if several mutational steps are needed to reach a sequence with sufficient binding energy to confer a selective advantage. As already pointed out by Berg *et al.* (2004) [32], this places strong constraints on the length of the binding sites, if they were to evolve from random sequences.

Several biophysical factors, such as TF concentration and the energetics of TF-DNA and TF-TF interactions, might play an important role in TFBS evolution. Quantitative models for TF sequence specificity [33–38] and for thermodynamic (TD) equilibrium of TF occupancy on DNA [34, 39–43] were developed in recent decades and, in parallel with developments in sequencing, have contributed to our understanding of TF-DNA interaction biophysics. These biophysical factors can shape the characteristics of the TFBS fitness landscape over genotype space in evolutionary models [8, 29, 32, 44–47]. There are also intensive efforts to understand the mapping from promoter/enhancer sequences to gene expression [42, 48–50]. Despite this recent attention, there have been relatively few attempts to understand the evolutionary dynamics of TFBS in full promoter/enhancer regions [29, 43, 51–53], especially using biophysically realistic but still mathematically tractable models. Such models are necessary to gain a thorough theoretical understanding of binding site evolution.

Our aim in this study is to investigate the dynamics of TFBS evolution by focusing on the typical evolutionary rates for individual TFBS gain and loss events. We consider both a single binding site at an isolated DNA region and a full enhancer/promoter region, able to harbor multiple binding sites. In the following section, we lay out our modeling framework, which covers both population genetic and biophysical considerations, as outlined above. Using this framework, we try to understand **i)** what typical gain and loss rates are for a single TFBS site; **ii)** how quickly populations converge to a stationary distribution for a single TFBS; **iii)** how multiple TFBS evolve in enhancers and promoters; **iv)** how early history of the evolving sequences can change the evolutionary rates of TFBS; and **v)** how cooperativity between TFs affects the evolution of gene expression. We find that, under realistic parameter ranges, both gain and loss of a single binding site is slow, slower than the typical divergence time between species. Importantly, fast emergence of an isolated TFBS requires strong selection and favorable initial sequences in the mutational neighborhood of a strong TFBS. The evolutionary process approaches the equilibrium distribution very slowly, raising concerns about the use of equilibrium assumptions in theoretical work. We proceed to show that the dynamics of TFBS evolution in larger sequences can be understood approximately from the dynamics of single binding sites; the TFBS gain times are again slow if evolution starts from random sequence in the absence of strong selection or large regulatory sequence "real estate." Finally, we identify two factors that can speed up the emergence of TFBS: the existence of an initial sequence distribution biased towards the mutational neighborhood of strongly binding sequences, which suggests that ancient evolutionary history can play a major role in the emergence of "novelties" [54]; and the biophysical cooperativity between transcription factors, which can partially account for the lack of observed correlation between identifiable binding sequences and transcriptional activity [11].

MODELS & METHODS

Population genetics

We consider a finite population of N diploid individuals whose genetic content consists of an evolvable L base pair (bp) contiguous regulatory sequence $\boldsymbol{\sigma}$ to which TFs can bind. Given that $\sigma_i \in \{A, C, G, T\}$ where $i = 1, 2, ..., L$ indexes the position in regulatory sequence, there are 4^L different regulatory sequences in the genotype space. Each TF is assumed to bind to a contiguous sequence of n bp within our focal region of L bp (Fig. 1A,B). Regulatory sequences evolve under mutation, selection, and sampling drift. The rest of the genome is assumed to be identical for all individuals and is kept constant. In the first part of our study we consider the regulatory sequence comprised of a single TFBS (i.e. $L = n$). Later, we consider the evolution of a longer sequence (i.e. $L \gg n$) in which more than one TFBS can evolve. For simulations, we use a Wright-Fisher model where N diploid individuals are sampled from the previous generation after mutation and selection. Our analytical treatment is general and corresponds to setups where a diffusion approximation to allele frequency evolution is valid. We neglect recombination since typical regulatory sequences are short, $L \leq 1000$. To be consistent with most of the population genetics literature we assume diploidy, but since we do not consider any dominance effects, our results also hold for a haploid population with $2N$ individuals.

Evolutionary dynamics simplify in the low mutation limit where the population consists of a single genotype during most of its evolutionary history (the fixed state population model). Desai & Fisher [55] have shown that the condition $\frac{\log 4N\Delta f}{\Delta f} \ll \frac{1}{4NU_b\Delta f}$ needs to hold for a fixed state population assumption to be accurate. The term on the left is the establishment time of a mutant allele with a selective advantage Δf relative to the wild type; the term on the right-hand side is the waiting time for such an allele to appear, where U_b is the beneficial mutation rate per individual per generation. Note that, in binding site context, U_b refers to the rate of mutations which increase the fitness, for instance, by increasing binding strength. Its exact value depends on the current state of the genotype; nevertheless, typical value estimates help model the evolutionary dynamics. In multicellular eukaryotes, where most evidence for the evolution of TFBSs has been collected and which provide the motivation for this manuscript, the number of mutations per nucleotide site is typically low, e.g. $4Nu \sim 0.01$ in *Drosophila* and $4Nu \sim 0.001$ in humans [56], where u is the point mutation rate per generation per base pair. For a single binding site of typical length $n \sim 5 - 15$, one therefore expects the fixed state population model to be accurate. For longer regulatory sequences, one expects that beneficial mutations are rare among all possible mutations, so that the fixed state population model can be assumed to hold as well.

Evolution under the fixed state assumption can be treated as a simple Markovian jump process. The transition rate from a regulatory sequence $\boldsymbol{\sigma}$ to another regulatory sequence $\boldsymbol{\sigma'}$ in a diploid population is

$$R_{\sigma',\sigma} = 2N\, U_{\sigma',\sigma}\, P_{\text{fix}}(N,\, \Delta f_{\sigma',\sigma}) \tag{1}$$

where $\Delta f_{\sigma',\sigma} = f(\boldsymbol{\sigma'}) - f(\boldsymbol{\sigma})$ is the fitness difference and $U_{\sigma',\sigma}$ is the mutation rate from $\boldsymbol{\sigma}$ to $\boldsymbol{\sigma'}$. The fixation probability P_{fix} of a mutation with fitness difference Δf in a diploid population of N individuals is

$$P_{\text{fix}}(N,\, \Delta f) = \frac{1 - e^{-2\Delta f}}{1 - e^{-4N\Delta f}} \approx \frac{2\Delta f}{1 - e^{-4N\Delta f}}, \tag{2}$$

which is based on the diffusion approximation [57]. Note that the fixation probability scaled with $1/N$ approximates to $2N\Delta f$ when $N\Delta f \gg 1$. Evolutionary dynamics therefore depend essentially on how regulatory sequences are mutationally connected in genotype space, and how fitnesses differ between neighboring genotypes, i.e., on the fitness landscape.

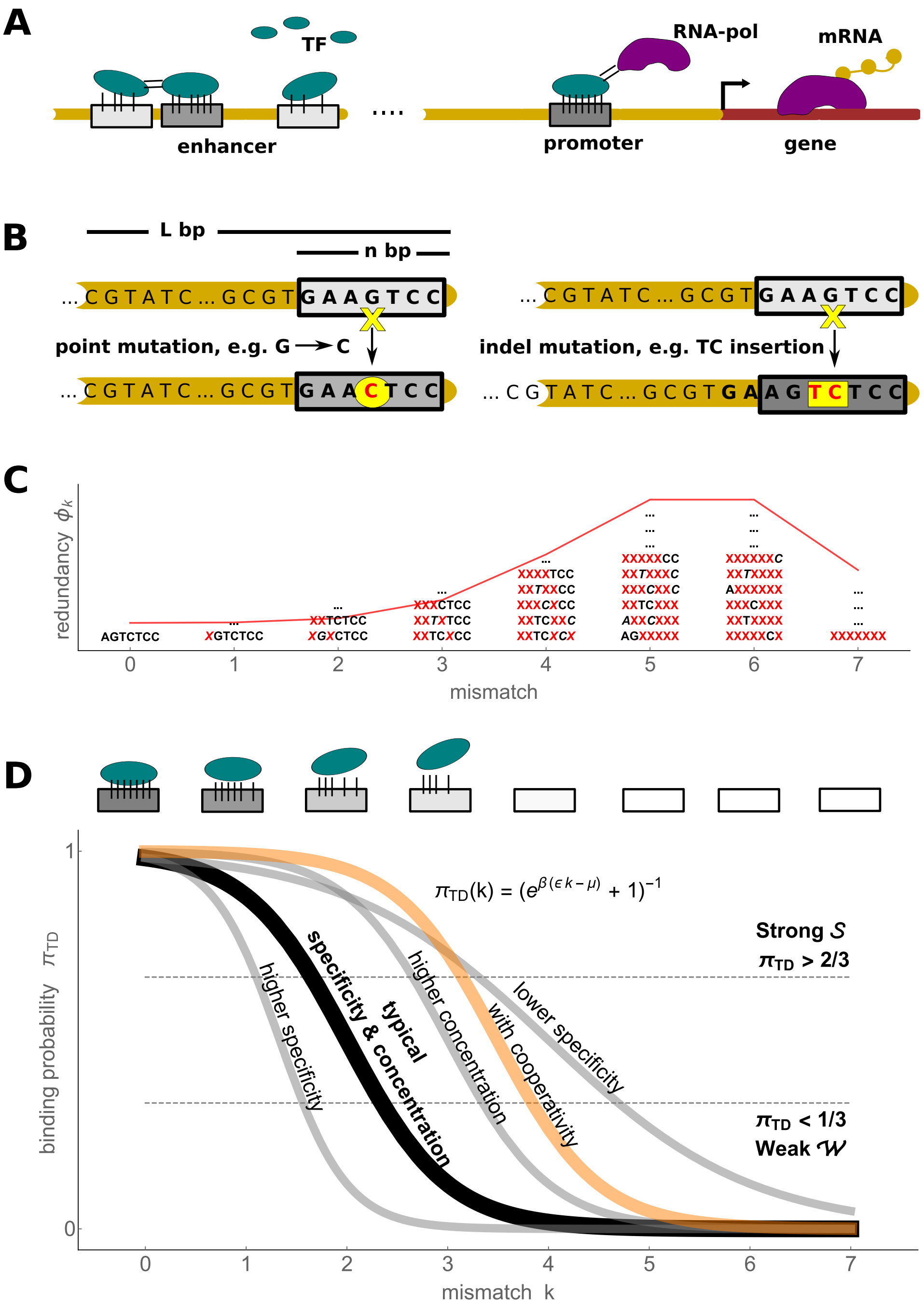

FIG. 1: **Biophysics of transcription regulation. A)** TFs bind to regulatory DNA regions (promoters and enhancers) in a sequence-specific manner to regulate transcriptional gene expression (mRNA production) level via different mechanisms, such as recruiting RNA polymerase (RNA-pol). **B)** A schematic of two types of mutational processes that we model: point mutations (left) and indel mutations (right). **C)** The mismatch binding model results in redundancy of genotype classes, with a binomial distribution (red) of genotypes in each mismatch class (some examples of degenerate sequences shown) **D)** The mapping from the TFBS regulatory sequence to gene expression level is determined by the thermodynamic occupancy (binding probability) of the binding site. If each of the k mismatches from the consensus sequence decreases the binding energy by ϵ, the occupancy of the binding site is $\pi_{\text{TD}}(k) = (1+e^{\beta(\epsilon k-\mu)})^{-1}$, where μ is the chemical potential (related to free TF concentration). A typical occupancy curve is shown in black ($\epsilon = 2\,k_BT$ and $\mu = 4\,k_BT$); the gray curves show the effect of perturbation to these parameters ($\epsilon = 1\,k_BT$, $\epsilon = 3\,k_BT$ and $\mu = 6\,k_BT$); the orange curve illustrates the case of two cooperatively binding TFs ($k_c = 0$ and $E_c = -3\,k_BT$, see text for details). We pick two thresholds, shown in dashed lines, to define discrete binding classes: strong $\mathcal{S}$ ($\pi_{\text{TD}} > 2/3$) and weak $\mathcal{W}$ ($\pi_{\text{TD}} < 1/3$).

Directional selection on biophysically motivated fitness landscapes

In this study, we focus on directional selection by assuming that fitness f is proportional to gene expression level g which depends on regulatory sequence, i.e.

$$f(\boldsymbol{\sigma}) = s\, g(\boldsymbol{\sigma}) \tag{3}$$

where s is the selection strength. It is important to note that this choice does not imply that directional selection is the only natural selection mechanism. It simply aims at obtaining the theoretical upper limits for the rates of gaining and losing binding sites.

To analyze a realistic but tractable mapping from the regulatory sequence to fitness, we primarily assume that the proxy for gene expression is the binding occupancy (binding probability) π at a single TF binding site, or the sum of the binding occupancies within an enhancer/promoter region (based on limited experimental support [84]). This corresponds to

$$f(\boldsymbol{\sigma}) = s \sum_i \pi^{(i)}(\boldsymbol{\sigma}) \tag{4}$$

where $\pi^{(i)}$ is the binding occupancy of a site starting at the nucleotide i in sequence $\boldsymbol{\sigma}$, and s can be interpreted as the selective advantage of a strongest binding to a weakest binding at a site. We assume all binding sites have equal strength and direction in their contribution towards total gene activation. Sites acting as repressors in our simple model would enter into Eq (4) with a negative selection strength, s. Future studies developing mathematically tractable models should consider more realistic case of unequal contribution with combined activator and repressor sites responding differentially to various regulatory inputs [53]. Although one can postulate different scenarios that map TF occupancies in a long ($L \gg n$) promoter to gene expression, we chose the simplest case which allows us to make analytical calculations. Later we relax our assumption on noninteracting binding sites and consider the effects of several kinds of interactions on gene expression and thus on evolutionary dynamics.

The occupancy of the TF on its binding site is assumed to be in thermodynamic (TD) equilibrium [34, 39–43]. While this might not always be realistic [58, 59], there is empirical support for this assumption (particularly in prokaryotes) [48, 60, 61], and more importantly, it is sufficient to capture the essential nonlinearity in this genotype-phenotype-fitness mapping [62]. In thermodynamic equilibrium, the binding occupancy at the site starting with the i-th position in regulatory sequence is given by

$$\pi_{\mathrm{TD}}^{(i)}(E_i) = \left(1 + e^{\beta(E_i-\mu)}\right)^{-1}. \tag{5}$$

Here, μ is the chemical potential of the TF (related to its free concentration) [44, 64]; E_i is the sequence specific binding energy, where lower energy corresponds to tighter binding, and $\beta = (k_B T)^{-1}$. We compute the binding energy E_i by adopting an additive energy model which is considered to be valid at least up to a few mismatches from the consensus sequence [37, 38, 65, 66], i.e.

$$E_i(\boldsymbol{\sigma}) = \sum_{j=i}^{i+n-1} \xi_{\sigma_j,j} \tag{6}$$

where ξ stands for the energy matrix whose $\xi_{\sigma_j,j}$ element gives the energetic contribution of the nucleotide σ_j appearing at the j-th position within TFBS. With this, Eq (4) can be rewritten more formally as

$$f(\boldsymbol{\sigma}) = s \sum_i \pi_{\mathrm{TD}}^{(i)}(E_i(\boldsymbol{\sigma})) \tag{7}$$

To allow analytical progress, we make the "mismatch assumption," i.e., the energy matrices contain identical $\epsilon > 0$ entries for every non-consensus (mismatch) base pair; the consensus entries are set to zero by convention. A single binding sequence with k mismatches therefore has the binding energy $E = k\epsilon$. We will refer to ϵ as "specificity." Specificity is provided by diverse interactions between DNA and TF, including specific hydrogen bonds, van der Waals forces, steric exclusions, unpaired polar atoms, etc. [63]. ϵ is expected to be in the range $1-3\ k_B T$, which is consistent with theoretical arguments [44] as well as direct measurements [65–67]. Note that we explicitly check the validity of the analytical results based on the mismatch assumption by comparing them against simulations using realistic energy matrices. The redundancy (i.e., normalized number of distinct sequences) of a mismatch class k at a single site in a random genome can be described by a binomial distribution $\boldsymbol{\phi}$ (Fig. 1C) where the probability of encountering a mismatch class k is

$$\boldsymbol{\phi}_{\boldsymbol{k}}(n, \alpha) = \binom{n}{k}\, \alpha^k \big(1-\alpha\big)^{n-k} \tag{8}$$

where $\alpha = 3/4$ in the case of equiprobable distribution over the four nucleotides.

We focus on selection in a single environment, which in this framework corresponds to a single choice for the TF concentration. We therefore fix the chemical potential to a baseline value of $\mu = 4\, k_B T$, which maps changes in the sequence (mismatch class k) to a full range of gene expression levels, as shown in Fig. 1D. We subsequently vary μ systematically and report how its value affects the results.

After these preliminaries, the equilibrium binding probability of Eq (5) reduces to

$$\pi_{\mathrm{TD}}(k) = \Big(1 + e^{\beta(\epsilon\, k - \mu)}\Big)^{-1}. \tag{9}$$

This function has a sigmoid shape whose steepness depends on specificity ϵ and whose midpoint depends on the ratio of chemical potential to specificity, μ/ϵ (Fig. 1D). To simplify discussion, we introduce two classes of sequences: genotypes are associated with "strong binding" $\mathcal{S}$ and "weak binding" $\mathcal{W}$ if $\pi_{\mathrm{TD}} > 2/3$ and $\pi_{\mathrm{TD}} < 1/3$, respectively. The thresholds that we pick are arbitrary, while still placing the midpoint of the sigmoid between the two classes; our results do not change qualitatively for other choices of thresholds. In the mismatch approximation, the genotype classes $k = \{0,\, 1,\, ...,\, k_{\mathcal{S}}\} \in \mathcal{S}$ and $k = \{k_{\mathcal{W}},\, k_{\mathcal{W}}+1,\, ...,\, n\} \in \mathcal{W}$ correspond to strong and weak binding, respectively. $k_{\mathcal{S}}$ and $k_{\mathcal{W}}$ are defined as the closest integers to the thresholds defined above; these values depend on ϵ and μ. We also define a "presite" as the mismatch class that is 1 bp away from the threshold for strong binding, i.e., a class with $k_{\mathcal{S}}+1$ mismatches. Note that binding length n extends the tail of the fitness landscape for a single site and shifts the center of redundancy rich mismatch classes (Fig. 1C).

The formulation in Eq (7) reduces to

$$f(k) = s\, \pi_{\mathrm{TD}}(k) \tag{10}$$

in a mismatch approximation at a single site, which we will investigate extensively for Ns scaling of TFBS gain and loss rates. We consider a wide range of Ns values: $Ns < 0$ for negative selection, $Ns = 0$ for neutral evolution, $Ns \sim 1$ for weak positive selection, $Ns \gg n \log(2)/2$ for strong positive selection (see below for this particular choice of the threshold).

In order to study the effects of interacting TFBSs in large regulatory sequences, we relax our assumption of non-interacting TFBS in Eq (7) and study three simple models. In the main text, we report the cooperative physical interaction between two TF molecules binding two nearby sites where the binding probability at a site is modified as

$$\pi_{\mathrm{coop}}(k,\, k_c) = \frac{e^{-\beta(\epsilon k-\mu)} + e^{-\beta(\epsilon(k+k_c)-2\mu-E_c)}}{1 + e^{-\beta(\epsilon k-\mu)} + e^{-\beta(\epsilon k_c-\mu)} + e^{-\beta(\epsilon(k+k_c)-2\mu-E_c)}}, \tag{11}$$

where k_c stands for the mismatch class at the co-binding site and E_c for cooperativity. In this study we consider that cooperative energy ranges from an intermediate strength ($E_c = -2\, k_B T$) to a high strength ($E_c = -4\, k_B T$) [42]. Fig 1D shows an example of the binding probability when a strong co-binding site exists. As a function of k alone, at fixed k_c, this formulation of cooperativity is consistent with the zero-cooperativity ($E_c = 0$) case but with a changed effective chemical potential. We take cooperative interactions into account if the two TFs are binding within 3 bp of each other, and we only consider the strongest binding of the cooperative partner (i.e., the proximal location with the lowest k_c).

In Supporting Information, we discuss the other two models of interacting TFBS. In one model, gene expression is determined only by the binding probability of the strongest site in the regulatory sequence. In the other model, gene expression is determined by the probability of the joint occupancy of 2 strongest binding sites, anywhere in the regulatory sequence; this model is a toy version of synergistic "non-physical" interaction of TFs which compete with nucleosomal binding for the occupancy of regulatory regions in eukaryotes (see Mirny (2010) [68] for a detailed model).

Point and indel mutations

Point mutations and indels are the only mutational processes in our framework. Point mutations with a rate u convert the nucleotide at one position into one of the 3 other nucleotide types. For a single binding site, the probability that a point mutation changes the mismatch class from k to k' is

$$\boldsymbol{P}^{(\text{point})}_{k',k} = \big(1 - k/n\big)\,\delta_{k',k+1} + \big(k/3n\big)\,\delta_{k',k-1} + \big(2k/3n\big)\,\delta_{k',k} \tag{12}$$

where $\delta_{a,b} = 1$ if $a = b$ and 0 otherwise.

We define the indel mutation rate per base pair such that it occurs with rate $\theta\, u$ at a position where a random nucleotide sequence is either inserted, or an existing nucleotide sequence is deleted. For mathematical simplicity, we assume that insertions and deletions are equally likely; in fact, a slight bias towards deletions is reported in the literature with a ratio of deletion to insertion $\sim 1.1 - 3.0$ [69–71]. Parameter θ is the ratio of indel mutation rate to point mutation rate, and is reported to be in the range $0.1 - 0.2$ [72–74]. We consider two cases: the baseline of $\theta = 0$ for no indel mutations, and $\theta = 0.15$ for the combined effect of indel and point mutations. Since we fix the length of the regulatory sequence, indels shift existing positions away from or inwards to some reference position (e.g., transcription start site). For consistency, we fix the regulatory sequence at its final position and assume that sequences before the initial position are random. Indel lengths vary, with reports suggesting a sharply decreasing but fat-tail frequency distribution [75]. For simulations we consider only very short indels of size $1 - 2$ bp, occurring proportional with their reported frequencies of 0.45 and 0.18, respectively. We do not need to assume any particular indel length for analytical calculations (below). While sufficient for our purposes, this setup would need to be modified when working with real sequence alignments of orthologous regions.

For a single binding site (i.e. $L = n$) one can exactly calculate the probability of an indel mutation changing the mismatch class from k to k' as

$$\boldsymbol{P}^{(\text{indel})}_{k',k} = \sum_{i=1}^{n} (1/n) \sum_{x=0}^{k'} p(X_i = x \mid k)\, p(Y_i = k' - x). \tag{13}$$

Here, i is the index for the position of an indel mutation within the binding site. The distribution over possible positions is uniform (hence $1/n$). The indel mutation defines two distinct parts in the binding site in terms of mismatches: nucleotides behind the indel mutation preserve their mismatch information, yet the nucleotides within and after indel mutation completely lose it. The new mismatches at these distinct parts X_i and Y_i are binomial random variables,

$$\begin{aligned} p(X_i = x \mid k) &= \boldsymbol{\phi_x}(i-1,\, \alpha = k/n) \\ p(Y_i = y) &= \boldsymbol{\phi_y}(n - i + 1,\, \alpha = 3/4) \end{aligned} \tag{14}$$

where $\boldsymbol{\phi_k}(n,\, \alpha)$ is defined in Eq (8). Fig 6 shows that Monte Carlo sampling of indel mutations at a single binding site matches the analytical expression in Eq (13).

The two types of mutations can be combined into the mutation rate matrix as follows:

$$\boldsymbol{U}_{k',k} = \begin{cases} n\ u\ \left(\boldsymbol{P}^{(\text{point})}_{k',k} + \theta\, \boldsymbol{P}^{(\text{indel})}_{k',k}\right) & k' \neq k \\ -\sum_{k' \neq k} \boldsymbol{U}_{k',k} & k' = k \end{cases}. \tag{15}$$

Evolutionary dynamics of single TF binding sites

For a sequence that consists of an isolated TFBS (i.e., $L = n$), analytical treatment is possible under the fixed state assumption. Let $\boldsymbol{\psi}(t)$ be a distribution over an ensemble of populations, whose k-th component, $\boldsymbol{\psi}_k(t)$, denotes the probability of detecting a genotype with k mismatches at time t. In the continuous time limit, the evolution of $\boldsymbol{\psi}(t)$ is described by

$$\frac{d}{dt}\boldsymbol{\psi}(t) = \boldsymbol{R} \cdot \boldsymbol{\psi} \tag{16}$$

which accepts the following solution:

$$\boldsymbol{\psi}(t) = e^{\boldsymbol{R}\, t} \cdot \boldsymbol{\psi}(0). \tag{17}$$

Here, $\boldsymbol{R}$ is the transition rate matrix defined as

$$\boldsymbol{R}_{k',k} = \begin{cases} 2N\, \boldsymbol{U}_{k',k}\ P_{\text{fix}}(N,\, \Delta f_{k',k}) & k' \neq k \\ -\sum_{k' \neq k} \boldsymbol{R}_{k',k} & k' = k \end{cases}. \tag{18}$$

This dynamical system is a continous-time Markov chain and there exists a unique stationary distribution $\hat{\boldsymbol{\psi}}$ corresponding the genotype distribution over an ensemble of populations at large time points. It can be calculated by decomposing the transition rate matrix $\boldsymbol{R}$ into its eigenvalues and eigenvectors. The normalised left eigenvector with zero eigenvalue corresponds to the stationary distribution. This can also be expressed analytically as

$$\hat{\boldsymbol{\psi}}_k \propto e^{F(k,N)+H(k\ \mid\ n,\alpha)}, \tag{19}$$

where $F(k,N) = 4Nf(k)$ captures the relative importance of selection to genetic drift, and $H(k \mid n,\alpha)$ is the mutational entropy, describing how a particular mismatch class k is favored due to redundancy and connectivity of the genotype space. For point mutations alone ($\theta = 0$), $H = \log\phi_k(n,\alpha)$, with the binomial distribution $\phi_k(n,\ \alpha)$ as defined in Eq (8). Obtaining a closed form expression for H is difficult when considering indel mutations ($\theta > 0$), yet the eigenvalue method solution suggests a similar shape for θ in the range of interest. The form of the stationary distribution was known for a long time in population genetics literature for a single locus or many loci with linkage equilibrium [76]. It has recently been generalised to arbitrary sequence space under the fixed state assumption [32, 77], resulting in the form of Eq (19) with a close analogy in the energy-entropy balance of statistical physics [80], and become a subject of theoretical interest [62, 78, 79, 81].

Under weak directional selection for high expression (and thus high binding site occupancy), the stationary distribution shows a bimodal shape, with one peak located around the fittest class, $k \sim 0$, and another at the core of mutational entropy, $k \sim \alpha\, n$ (recall that $\alpha = 3/4$ for a completely random genome). This bimodal shape collapses to a unimodal one, either at no selection or at strong selection. The threshold value for Ns distinguishing strong and weak selection regimes primarily depends on the TFBS binding length, n. In a sigmoidal fitness landscape and approximating the binomial distribution by a normal distribution as appropriate, the sizes of these two peaks are roughly proportional to $\exp\left(4Ns - n\log 4\right)$ and $\sqrt{2\pi\alpha(1-\alpha)n}$, respectively. Therefore, we expect the threshold Ns to scale as $\frac{1}{4}\Big(n\,\log 4 - \frac{1}{2}\log 2\pi\alpha(1-\alpha)n\Big)$. For typical n, the linear term is dominant, suggesting that

$$Ns \sim n\log(2)/2 \tag{20}$$

corresponds to the threshold for strong selection in TFBS evolution (cf. Fig 7). Note that this n scaling differs from the $\log(n)$ scaling which is expected in simple fitness landscapes [82]. Our argument assumes that the system is at evolutionary equilibrium, which, as we will see, is not necessarily the case even under strong selection, providing further motivation for focusing on dynamical aspects of evolution.

We define the time needed to gain (or lose) a TFBS as the time it takes for a strong binding site to emerge from a weak one (and vice versa), as schematized in Fig. 1D. For an isolated TFBS, these times can be computed from the Markovian properties of the evolutionary dynamics, by calculating the average first hitting times [83]. We will use the notations $\langle t\rangle_{\mathcal{S}\leftarrow k}$ and $\langle t\rangle_{\mathcal{W}\leftarrow k}$, respectively, for average gain and loss times when evolution starts from mismatch class k. Obviously, $\langle t\rangle_{\mathcal{S}\leftarrow k} = 0$ if k is among the strong binding classes ($k \in \mathcal{S}$) and $\langle t\rangle_{\mathcal{W}\leftarrow k} = 0$ if k is among the weak binding classes ($k \in \mathcal{W}$). The average gain times from other mismatch classes can be found by considering the relation $\langle t\rangle_{\mathcal{S}\leftarrow k} = 1 + \sum_{k'\notin\mathcal{S}} \boldsymbol{P}_{k,k'}\langle t\rangle_{\mathcal{S}\leftarrow k'}$, where $\boldsymbol{P}_{k,k'}$ is the probability of transition from k' to k in one generation. One can compute the average gain times by writing it in terms of linear algebraic equation:

$$\boldsymbol{T}_{\mathcal{S}\leftarrow} = \left(\mathbf{R}_{\notin\mathcal{S}}\right)^{-\mathrm{T}} \cdot (-\mathbf{1}) \tag{21}$$

where $\boldsymbol{T}_{\mathcal{S}\leftarrow}$ is a column vector listing non-trivial gain times, i.e. $\{\langle t\rangle_{S\leftarrow k}\}$ for $k = k_{\mathcal{S}}+1,\ ...,\ n$. $\mathbf{R}_{\notin\mathcal{S}}$ is the $\mathbf{R}$ matrix with all rows and columns corresponding to $k \in \mathcal{S}$ deleted and $-\mathrm{T}$ is the matrix operator for the transpose after an inverse operation. $\mathbf{1}$ is a vector of ones. Similarly one can find the loss times,

$$\boldsymbol{T}_{\mathcal{W}\leftarrow} = \left(\mathbf{R}_{\notin\mathcal{W}}\right)^{-\mathrm{T}} \cdot (-\mathbf{1}) \tag{22}$$

where $\boldsymbol{T}_{\mathcal{W}\leftarrow}$ is a column vector listing non-trivial loss times, i.e. $\{\langle t\rangle_{\mathcal{W}\leftarrow k}\}$ for $k = 1,\ 2,\ ...\ k_W - 1$. $\mathbf{R}_{\notin\mathcal{W}}$ is the $\mathbf{R}$ matrix with all rows and columns corresponding to $k \in \mathcal{W}$ deleted.

In the case of point mutations alone ($\theta = 0$), the $\mathbf{R}$ matrix is tri-diagonal and one can deduce simpler formulae for gain and loss times:

$$\begin{aligned} \langle t\rangle^{(\mathrm{point})}_{\mathcal{S}\leftarrow k} &= \textstyle\sum_{i=k_{\mathcal{S}}+1}^{k} \frac{1}{\boldsymbol{R}_{i-1,\,i}}\,\frac{1-\hat{\boldsymbol{\Psi}}_{i-1}}{\hat{\boldsymbol{\psi}}_i} \\ \langle t\rangle^{(\mathrm{point})}_{\mathcal{W}\leftarrow k} &= \textstyle\sum_{i=k+1}^{k_{\mathcal{W}}} \frac{1}{\boldsymbol{R}_{i-1,\,i}}\,\frac{\hat{\boldsymbol{\Psi}}_{i-1}}{\hat{\boldsymbol{\psi}}_i} \end{aligned} \tag{23}$$

where we use $\hat{\boldsymbol{\Psi}}_i = \sum_{j=0}^{i} \hat{\boldsymbol{\psi}}_j$ to denote the cumulative stationary distribution. For very strong selection, the second term in the sums approaches unity, resulting in even simpler formulae [32], called the "shortest path" (sp) solution:

$$\begin{aligned} \langle t \rangle^{(\mathrm{sp})}_{\mathcal{S}\leftarrow k} &= \textstyle\sum_{i=k_{\mathcal{S}}+1}^{k} \frac{1}{\boldsymbol{R}_{i-1,\,i}} \\ \langle t \rangle^{(\mathrm{sp})}_{\mathcal{W}\leftarrow k} &= \textstyle\sum_{i=k+1}^{k_{\mathcal{W}}} \frac{1}{\boldsymbol{R}_{i-1,\,i}} \end{aligned} \quad . \tag{24}$$

These equations can be used to quickly estimate gain and loss rates of interest. For example, the gain rate from presites under strong selection is approximately $2\, Ns\, u \frac{k_{\mathcal{S}}+1}{3}(f(k_{\mathcal{S}}) - f(k_{\mathcal{S}}+1))$. Although the exact value depends on the binding specificity and chemical potential, one can see that it is about $Ns\, u$ for the parameter range of interest. Similarly, one can see that the rate of loss from strong sites is about $2n\, |Ns|\, u$ when there is strong negative selection.

RESULTS

Single TF binding site gain and loss rates under mutation-selection-drift are typically slow

We first studied the evolutionary rates for a single TF binding site at an isolated DNA sequence of the same length under mutation, genetic drift, and directional selection for high gene expression level (i.e., tighter binding). As detailed in the Models & Methods section, we combined a thermodynamically motivated fitness landscape with the mismatch approximation, and assumed that the mutation rate is low enough for the fixed state population approximation to be valid. Under these assumptions, we could calculate the inverse of the average TFBS gain and loss times as a function of the starting genotype, using either an exact method or Wright-Fisher simulations. We considered point mutations alone, or point mutations combined with short indel mutations, in order to understand under which conditions the rates of gaining and losing binding sites can reach or exceed the rates $2-3$ orders of magnitude greater than point mutation rate, and thus to become comparable to rates observed in comparative genomic studies.

Fig. 2A shows the dependence of the TFBS gain rate on the selection strength (with respect to genetic drift), Ns. For parameters typical of eukaryotic binding sites (length $n = 7$ bp, specificity $\epsilon = 2\, k_B T$), the TFBS gain rates are extremely slow (practically no evolution) when there is negligible selection pressure ($Ns \sim 0$), indicating the importance of selection for TFBS emergence. Indeed, the effective selection needs to be very strong, e.g., $Ns > 100$, for TFBS evolution to exceed the per-nucleotide mutation rate by orders of magnitude and become comparable to speciation rates.

Even if strong selection were present, the gain rate depends crucially on the initial genotype. While gain rates from presites, i.e., genotypes one mutation away from the threshold for strong binding, are roughly $Ns\, u$ for the strong Ns regime (as estimated by Berg *et al.* [32]), they decrease dramatically if more mutational steps are needed to evolve a functionally strong binding site. This is illustrated in the inset to Fig 2A, showing an exponential-like decay in the gain rates as a function of the number of mismatches, even for a TFBS of a modest length of 7 bp. As argued in the Models & Methods section (see Eq (20)), we confirmed that the threshold for the strong Ns regime scales as $n \log(2)/2$ and not as $\log(n)$ which is the case for simple fitness landscapes [82].

The availability of a realistic fraction of indel mutations (here, $\theta = 0.15$) can speed up evolution when starting from distant genotypes (cf. solid and dashed red line in Fig 2A). This is because indels connect the genotype space such that paths from many to few mismatches are possible within a single mutational step. Nevertheless, the improvement due to indel mutations does not alleviate the need for very strong selective pressure and the proximity of the initial to strongly-binding sequence, in order to evolve a functional site.

Biophysical parameters—the binding site length n, the chemical potential μ, and the specificity ϵ—influence the shape of the fitness landscape and thus the TFBS gain rates. This is especially evident when we consider *de novo* evolution starting from random sequence. As shown in Figs 2B, C, increases in specificity or length cause a sharp drop in the gain rates from initial sequences in the most redundancy rich class, which can be only partially mitigated by the availability of indel mutations. This especially suggests that adaptation of TFBS from random sequences for TF with very large binding lengths and very strong specificities is unlikely with point and indel mutations which can constrain the evolution of TF lengths and TF specificity, which is consistent with Berg *et al.* (2004) [32]'s earlier numeric observation. Importantly, the binding specificity and length show an inverse relation with the logarithm of the gain rates. This is due to the fact that a decrease in specificity allows more genotypes to generate appreciable binding and therefore fitness (see Fig. 1D), which partially compensates the increase in mutational entropy at larger binding site lengths. Variation of the chemical potential μ corresponding to an order-of-magnitude change in the free TF concentration does not qualitatively affect the results.

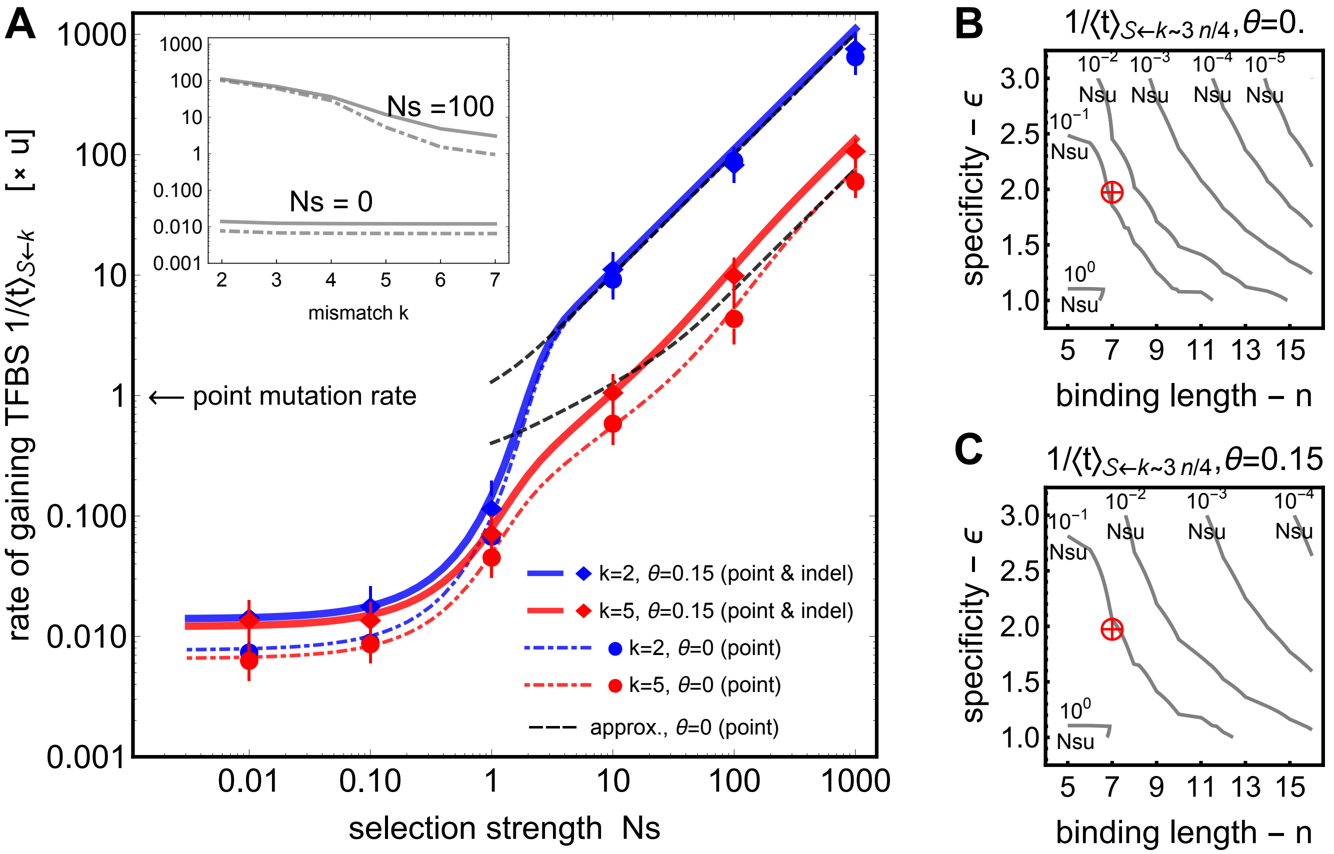

FIG. 2: **Single TF binding site gain rates at an isolated DNA region. A)** The dependence of the gain rate, $1/\langle t\rangle_{\mathcal{S}\leftarrow k}$ shown in units of point mutation rate, from sequences in different initial mismatch classes k (blue: $k = 2$, red: $k = 5$), as a function of selection strength. Results with point mutations only ($\theta = 0$) are shown by dashed line; with admixture of indel mutations ($\theta = 0.15$) by a solid line. For strong selection, $Ns \gg n\log(2)/2$, the rates scale with Ns, which is captured well by the "shortest path" approximation (black dashed lines in the main figure) of Eq (24). The biophysical parameters are: site length $n = 7$ bp; binding specificity $\epsilon = 2\ k_BT$; chemical potential $\mu = 4\ k_BT$. Points correspond to Wright-Fisher simulations with $Nu = 0.01$ where error bars cover ±2 SEM (standard error of mean). Inset shows the behavior of the gain rates as a function of the initial mismatch class k for $Ns = 0$ and $Ns = 100$. **B, C)** Gain rates from redundancy rich classes ($k \sim 3n/4$, typical of evolution from random "virgin" sequence) under strong selection, without (B) and with (C) indel mutations supplementing the point mutations. Red crosshairs denote the cases depicted in panel A. Contour lines show constant gain rates in units of Nsu as a function of biophysical parameters n and ϵ. Wiggles in the contour lines are not a numerical artefact but a consequence of discrete mismatch classes.

Typically slow TFBS evolution is a consequence of the sigmoidal shape of the thermodynamically motivated fitness landscape, where adaptive evolution in the redundant but weakly binding classes $\mathcal{W}$ must proceed very slowly due to the absence of a selection gradient. To illustrate this point, we generated alternative fitness landscapes that agree exactly with the thermodynamically motivated one from the fittest class to the threshold class for strong binding, $k_{\mathcal{S}}$, but after that decay as power laws, $\pi_{\rm pl}$, with a tunable exponent (see SI text). As seen in Fig 8, this exponent is a major determinant of the gain rates, suggesting that a biophysically realistic fitness landscape is crucial for the quantitative understanding of TFBS evolution.

To check that the assumption of the fixed state population is valid at $Nu = 0.01$, the value used here that is also relevant for multicellular eukaryotes [56], we performed Wright-Fisher simulations as described in the Models & Methods section. Fig 2A shows excellent agreement between the analytical results and the simulation. We further increased the mutation rate to $Nu = 0.1$, a regime more relevant for prokaryotes where polymorphisms in the population are no longer negligible, to find that the analytical fixed state assumption systematically overestimates the gain rates, as shown in Fig 9. In the presence of polymorphism, therefore, evolution at best proceeds as quickly as in monomorphic populations, and generally proceeds slower, so that our results provide a theoretical bound on the speed of adaptive evolution under directional selection. This is expected since the effects of clonal interference kick in after a certain Nu, where two different beneficial mutants start competing with each other, and eventually decrease the fixation probability in comparison to one beneficial mutant sweeping to fixation as in the monomorphic population case.

To check that the mismatch assumption does not strongly affect the reported results, we analyzed evolutionary dynamics with more realistic models of TF-DNA interaction. Different positions within the binding site can have different specificities, and one could suspect that this can significantly lower the evolutionary times. First, some positions within the TFBS may show almost no specificity for any nucleotide, most likely due to the geometry of TF-DNA interactions (e.g, when the TF can contact the nucleic acid residues only in the major groove); we have not simulated such cases explicitly, but simply take the binding site length n to be the effective sequence length where TF does make specific contacts with the DNA. Second, the positions that do exhibit specificity might do so in a manner that is more inhomogeneous than our mismatch assumption, which assigns zero energy to the consensus and a constant ϵ to any possible mismatch. We thus generated energy matrices where ϵ was drawn from a Gaussian distribution with the same mean $\langle\epsilon\rangle = 2\,k_BT$ as in our baseline case of Fig 2A, but with a standard deviation $0.5\,k_BT$. Fig 10 shows that both equal and unequal energy contributions produce statistically similar behaviors, indicating that inhomogeneous binding interactions cannot substantially enhance the evolutionary rates.

We further investigated the rate of TFBS loss (Fig 11). Here too strong (negative) selection is needed to lose a site on reasonable timescales, and it is highly unlikely that a site would be lost in the presence of positive selection. In contrast to the TFBS gain case, however, negative selection and mutational entropy act in the same direction for TFBS loss, reducing the importance of the initial genotype and making selection more effective at larger n and ϵ.

Taken together, these results suggest that the emergence of an isolated TFBS under weak or no selection is typically slow relative to the species' divergence times, and gets rapidly slower for sites that are either longer or whose TFs are more specific than the baseline case considered here. This suggests that biophysical parameters themselves may be under evolutionary constraints; in particular, if point mutations and indels were the only mutational mechanisms, the evolution of long sites, e.g. $n \gg 10-12$, would seem extremely unlikely, as has been pointed out previously [32]. Absent any mechanisms that could lead to faster evolution and which we consider below, isolated TFBS are generally only likely to emerge in the presence of strong directional selection and a favorable distribution of initial sequences that is enriched in presites.

Convergence to the stationary distribution is slow and depends strongly on initial conditions

A number of previous studies (e.g., [62, 78, 79]) assumed that a stationary distribution of mismatch classes is reached in the evolution of isolated TFBS and thus an equilibrium solution, Eq (19), is informative for binding sequence distributions. In contrast, our results for average gain and loss times suggest that the evolution of an isolated TFBS is typically slow. To analyze this problem in a way that does not depend on arbitrary thresholds defining "strong" and "weak" binding classes $\mathcal{S}$ and $\mathcal{W}$, we first examined the evolution of the distribution $\boldsymbol{\psi}(k)$ over the mismatch classes as a function of time in Fig 3A. For typical parameter values it takes on the order of the inverse point mutation rate to reach the stationary distribution for populations that start off far away from it, even with strong selection.

A systematic study of the convergence rates can be performed by computing the (absolute value of the) second eigenvalue, $|\lambda_2|$, of the transition rate matrix $\mathbf{R}$ from Eq (18), and exploring how this depends on the biophysical parameters n and ϵ. Consistent with previous results, we observe large increases in convergence times as n and ϵ increase. For example, an increase in the binding site length from $n = 7$ to $n = 11$ at baseline specificity of $\epsilon = 2\,k_BT$ would result in a ten-fold increase in the convergence time.

The intuitive reason behind the slow convergence rates is in the bimodal nature of the distribution $\boldsymbol{\psi}(k)$ on the thermodynamically motivated fitness landscape, similar to that reported by Lynch & Hagner [9]. One "attractor" is located around the fittest class ($k \sim 0$, due to directional selection), while the other is located around the redundancy-rich mismatch classes ($k \sim 3/4n$). These two attractors are separated by a typically sharp fitness landscape, and the redundancy-rich attractor lacks selection gradients needed to support fast adaptation. The temporal evolution of the distribution $\boldsymbol{\psi}(k)$ from, e.g., a maladapted state, can thus be best understood as the probability weight "switching" from resting approximately within one attractor to the other one, while maintaining the bimodal shape throughout, rather than a gradual shift of a unimodal distribution from a maladapted initial value of k to the value favored by selection. This is especially true when n gets larger: although adaptation within the functional sites can still happen, adaptation from the most random mismatch classes becomes extremely slow, even under strong selection (see Fig 15).

These results suggest that stationary distributions of isolated TFBS sequences may not be realizable on the timescales of speciation, which should be a cause of concern when stationarity is assumed without prior critical

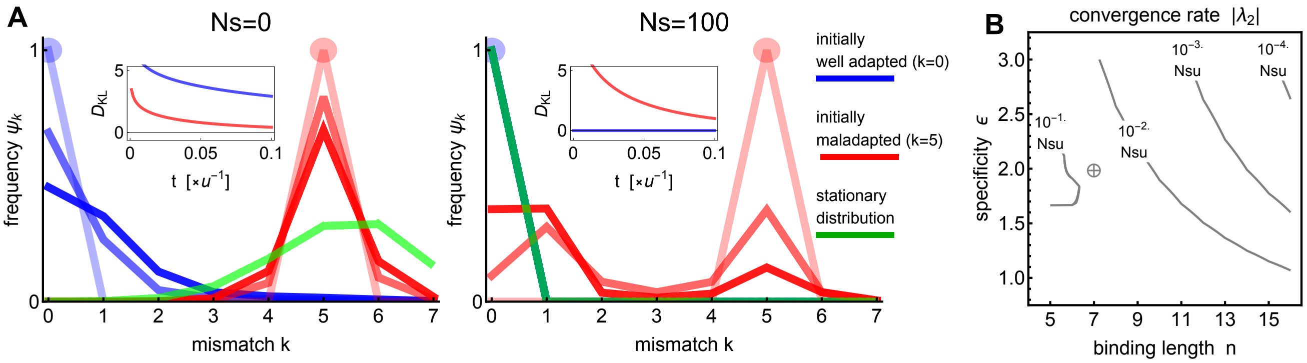

FIG. 3: **Convergence to the stationary distribution of TFBS sequences. A)** Evolutionary dynamics of the mismatch classes distribution $\boldsymbol{\psi}(k)$ for an isolated TFBS under point and indel mutations ($\theta = 0.15$), directional selection for stronger binding, and genetic drift is shown for initially well ($k = 0$, blue) and badly ($k = 5$, red) adapted populations. At left, no selection ($Ns = 0$); at right, strong selection ($Ns = 100$). Different curves show the distribution of genotype classes at different time points ($t = 0u^{-1}$, $0.05u^{-1}$, $0.1u^{-1}$ as decreasing opacity); stationary distribution is shown in green. Insets show the time evolution to convergence for initially well ($k = 0$, blue) and badly ($k = 5$, red) adapted populations, measured by the Kullback-Leibler divergence $D_{KL}[\boldsymbol{\psi}(t)\,||\,\boldsymbol{\psi}(t = \infty)]$. The biophysical parameters are: $n = 7$ bp, $\epsilon = 2\ k_BT$, $\mu = 4\ k_BT$. **B)** Rate of convergence to the stationary distribution for different ϵ and n values under strong selection ($Ns \gg n\log(2)/2$; here specifically $Ns = 100$) and for $\theta = 0.15$. Crosshairs represent the parameters used in a).

assessment. For example, applications assuming the stationary distribution might wrongly infer selection on regulatory DNA.

Evolution of TF binding sites in longer sequences

So far we have shown that the evolution of isolated TFBS is typically slow. How do the results change if we consider TFBS evolution in a stretch of sequence L bp in length, where $L \gg n$, e.g., within a promoter or enhancer? Here we focus on *de novo* evolution under strong directional selection for high gene expression, by simulating the process in the fixed state population framework. Compared to the isolated TFBS case, we need to make one further assumption: that the expression level of the selected gene is proportional to the summed TF occupancy on all sites within the regulatory region of length L (see Models & Methods for details). While this is the simplest choice, it is neither unique nor perhaps the most biologically plausible one, although limited experimental support exists for such additivity [84]; it does, however, represent a tractable starting point when the interactions between individual TF binding sites are not strong and the contribution of each site is equal and of the same sign. To address the interactions, we look at the cooperative binding case in the following section. In Supporting Information, we also discuss the competition of TFBSs for the strongest binding, and the "nonphysical" synergetic interaction by two strongest TFBSs.

We propose a simple analytical model for the time evolution of the number of strongly binding sites, $z(t)$, in the promoter, derived from isolated TFBS gain and loss rates, λ_{gain} and λ_{loss}. Assuming constant rates, one can write

$$\frac{d}{dt}z(t) = \lambda_{\text{gain}}\Big(z_{\max} - z(t)\Big) - \lambda_{\text{loss}}z(t) \tag{25}$$

where $z_{\max}$ is the maximum number of TFBS that can fit into the regulatory sequence of length L bp. If the sites can overlap, $z_{\max} = L - n + 1$, otherwise $z_{\max} \approx L/n$. The solution for Eq (25) is

$$z(t) = \Big(z_{\text{o}} - \frac{B}{A}\Big)e^{-At} + \frac{B}{A} \tag{26}$$

where $A = \big(\lambda_{\text{gain}} + \lambda_{\text{loss}}\big)$, $B = z_{\max}\lambda_{\text{gain}}$ and $z_{\text{o}} = z(t = 0)$. Under strong positive selection, i.e. $Ns \gg n\log(2)/2$, the loss rate λ_{loss} can be ignored. If the distribution of the initial mismatch classes in the promoter is ψ_k, one can approximate $z_{\max} - z_{\text{o}} = z_{\max}\sum_{k=k_{\mathcal{S}}+1}^{n}\psi_k$ to obtain:

$$z(t) - z_{\text{o}} = \big(1 - e^{-\lambda_{\text{gain}}t}\big)\, z_{\max} \sum_{k=k_{\mathcal{S}}+1}^{n} \psi_k. \tag{27}$$

There are two limiting regimes in which we can examine the behavior of Eq (27). Over a short timescale, evolutionary dynamics will search over all possible positions, $z_{\max} = L - n + 1$, to pull out the presites, since they are fastest to evolve into the strong binding class $\mathcal{S}$, i.e.:

$$\lambda_{\text{gain}} \approx \lambda_{\text{gain}}^{\text{presite}} = \Big(\sum_{k \notin \mathcal{S}} \psi_k\Big)^{-1} \psi_{k_{\mathcal{S}}+1} / \langle t \rangle_{\mathcal{S} \leftarrow k_{\mathcal{S}}+1} \tag{28}$$

As the process unfolds and new sites are established, new TFBS will only be able to emerge at a smaller set of positions due to possible overlaps, so that $z_{\max} \approx L/n$. On the other hand, evolution from higher mismatch classes will also start to contribute towards new sites:

$$\lambda_{\text{gain}} \approx \lambda_{\text{gain}}^{\text{all}} = \Big(\sum_{k \notin \mathcal{S}} \psi_k\Big)^{-1} \sum_{k \notin \mathcal{S}} \psi_k / \langle t \rangle_{\mathcal{S} \leftarrow k} \tag{29}$$

Fig 4 shows how new TFBSs with length $n = 7$ bp emerge over time in a promoter of $L = 30$ bp in length. Consistent with the predictions of our simplified model, we can distinguish the early, intermediate, and late epochs. In the early epoch, $t < 0.01u^{-1}$, presites are localized among all possible locations and are established as binding sites. During this period, the growth in the expected number of new TFBSs is linear with time. The importance and predictive power of presites at early epoch remain even under different models of gene expression, including interaction between TFBSs (see Fig 14). In the intermediate epoch, new binding sites accumulate at the rate that is slightly above that expected by establishment from presites alone, as the mutational neighborhood is explored further. In the late epoch, $t > 0.1u^{-1}$, initial sites in the immediate mutational vicinity have been exhausted, and established sites have constrained the number of positions where new sites can evolve from more distant initial sequences, leading to the saturation in the number of evolved TFBS.

Using the simple analytical model, we explored in Fig 4B,C how the binding length n and specificity ϵ affect the number of newly evolved TFBS. Increasing n leads to a steep decrease in the number of expected sites, with a somewhat weaker dependence on ϵ, especially at early times. Simulations at other values of biophysical and evolutionary parameters confirm the qualitative agreement between the analytical model and the simulation (Fig 12); given that the model is a simple heuristic, it cannot be expected to match the simulations in detail, yet it nevertheless seems to capture the gross features of evolutionary dynamics. Together, these results show that at early times under strong selection, the number of newly evolved sites will grow linearly with time and proportional to L, before evolution from higher mismatch classes can contribute and ultimately before the sites start interacting, with a consequent slowdown in their evolution. Thus, evolution in longer regulatory regions ($L = 10^2 - 10^3$ bp) could feasibly give rise to tens of binding sites at $Ns = 10^2 - 10^3$ within a realistic time frame $t \sim 0.001u^{-1}$, if the sites are sufficiently short ($n \sim 7$ bp). Explaining the evolution of longer sites, e.g., $n > 10 - 12$ bp, especially within short promoters found in prokaryotes, would likely necessitate invoking new mechanisms.

Ancient sites and cooperativity between TFs can accelerate binding site emergence

Finally, we briefly examine two mechanisms that can further speed up the evolution of TF binding sites in longer sequences.

The first possibility is that the sequence from which new TFBS evolve is not truly random; as discussed previously, presites have a strong influence on the early accumulation of new binding sites. There are a number of mechanisms that could bias the initial sequence distribution towards presites: examples include transposable elements, DNA repeats, or CG content bias. Here we consider an alternative mechanism that we refer to as the "ancient TFBS scenario," in which a strong TFBS existed in the sequence in the ancient past, after which it decayed into a weak binding site, possibly due to the relaxation of selection (i.e., $Ns \sim 0$).

As we demonstrated in the context of isolated sites, TFBS loss rates are slow and the remains of the binding site will linger in the sequence for a long time before decaying into the most redundancy rich mismatch classes. This biased initial distribution of mismatches $\boldsymbol{\Psi}$ in a sequence of length L with a single ancient site can be captured by writing:

$$\boldsymbol{\Psi} = \frac{1}{L-n+1}\, \boldsymbol{\psi}(t') + \frac{L-n}{L-n+1}\, \boldsymbol{\phi} \tag{30}$$

where $\boldsymbol{\phi}$ is the binomial distribution, Eq (8), characteristic of the random background, and $\boldsymbol{\psi}(t')$ is the distribution of mismatches due to the presence of the ancient site. Time t' refers to the interval in which the isolated ancient TFBS has been decaying under relaxed selection, and the corresponding $\boldsymbol{\psi}(t')$ can be solved for using Eq (17).

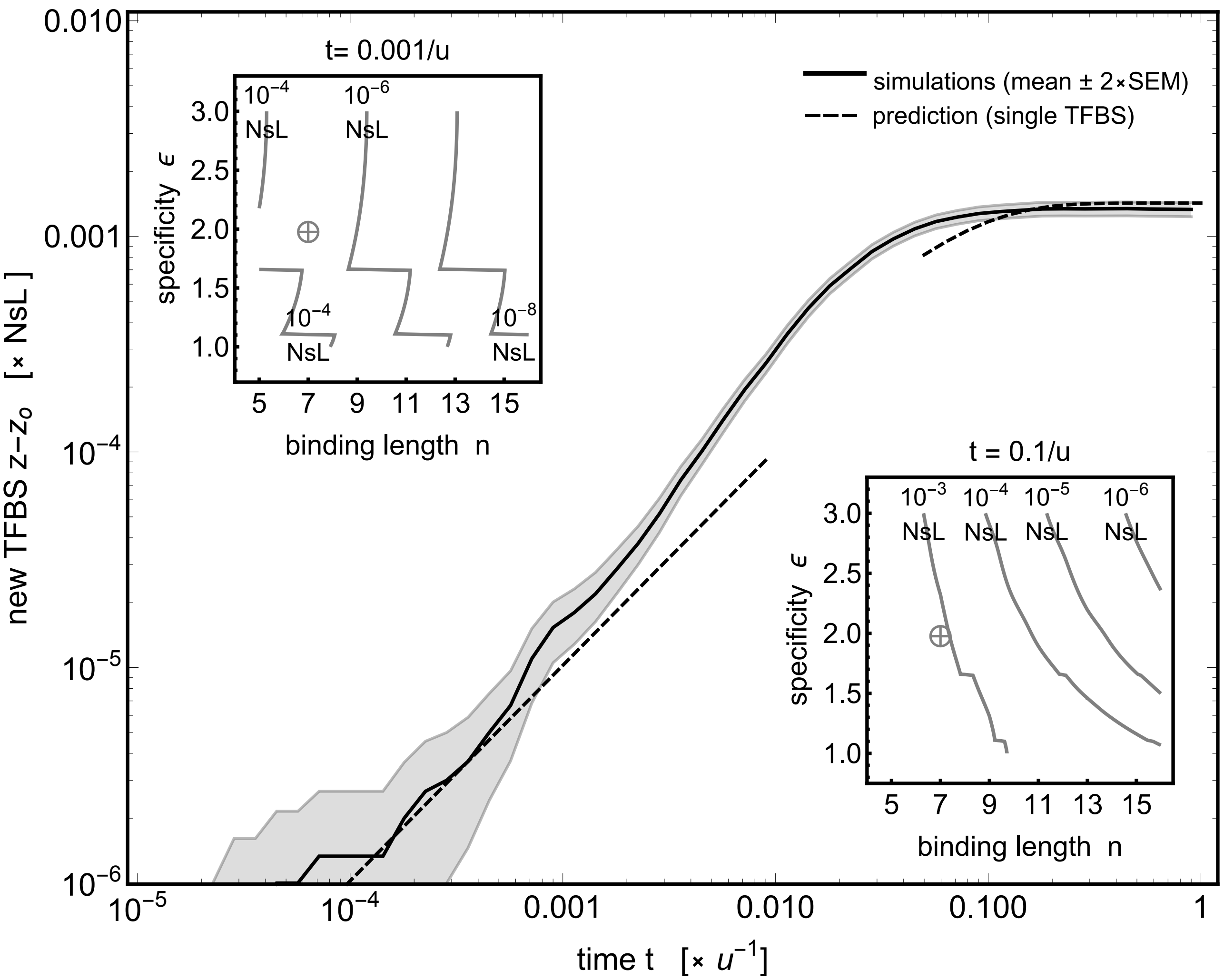

FIG. 4: **TF binding site evolution in a longer sequence of $L = 30$ base pairs.** The expected number of newly evolved TF binding sites with length $n = 7$ bp, under strong directional selection ($Ns = 100$) and both point and indel mutations ($\theta = 0.15$). Time is measured in inverse mutation rates; the number of newly evolved sites is scaled to the selection strength and the sequence length. 1000 replicate simulations were performed with different initial sequences. Average number of sites shown by a solid black line; the gray band shows ±2 SEM (standard error of the mean) envelope. Dashed curves are analytical predictions based on single TFBS gain rates at an isolated DNA region, given by Eqs (27,28,29). Biophysical parameters used: $\epsilon = 2\ k_BT$, $\mu = 4\ k_BT$. **Insets:** Expected number of newly evolved sites from a random sequence of length L at $t = 0.001u^{-1}$ (left) and $t = 0.1u^{-1}$ (right) for different binding length and specificity values, computed using the analytical predictions. Crosshairs denote the values used in the main panel.

Fig 5A shows that the ancient site scenario can enhance the number of newly evolved sites by resurrecting the ancient site, even after it has decayed for $t' = 0.1u^{-1}$. Simulation results agree well with the simple analytical model using the biased initial sequence distribution of Eq (30). Importantly, such a mechanism is particularly effective for longer binding sites of high specificity, indicating that regulatory sequence reuse could be evolutionarily beneficial in this biophysical regime (see Fig 13).

Fig 5A and Fig 13 also show the emergence of new sites when the ancient site was not a full consensus (preferred) sequence but differed from it by a certain number of mismatches. The results qualitatively agree with the case of perfect consensus. Importantly, this shows that the applicability of the ancient site scenario extends to cases where the ancient site belonged to a different TF (albeit with a preferred sequence similar to the studied TF), which has recently been reported to be a frequent phenomenon by Payne & Wagner (2014) [47], possibly due to evolution of TFs by duplication and divergence [85].

The second mechanism that we consider is the physical cooperativity between TFs: when one site is occupied, it is favorable for the nearby site to be occupied as well. We extended the thermodynamic model to incorporate

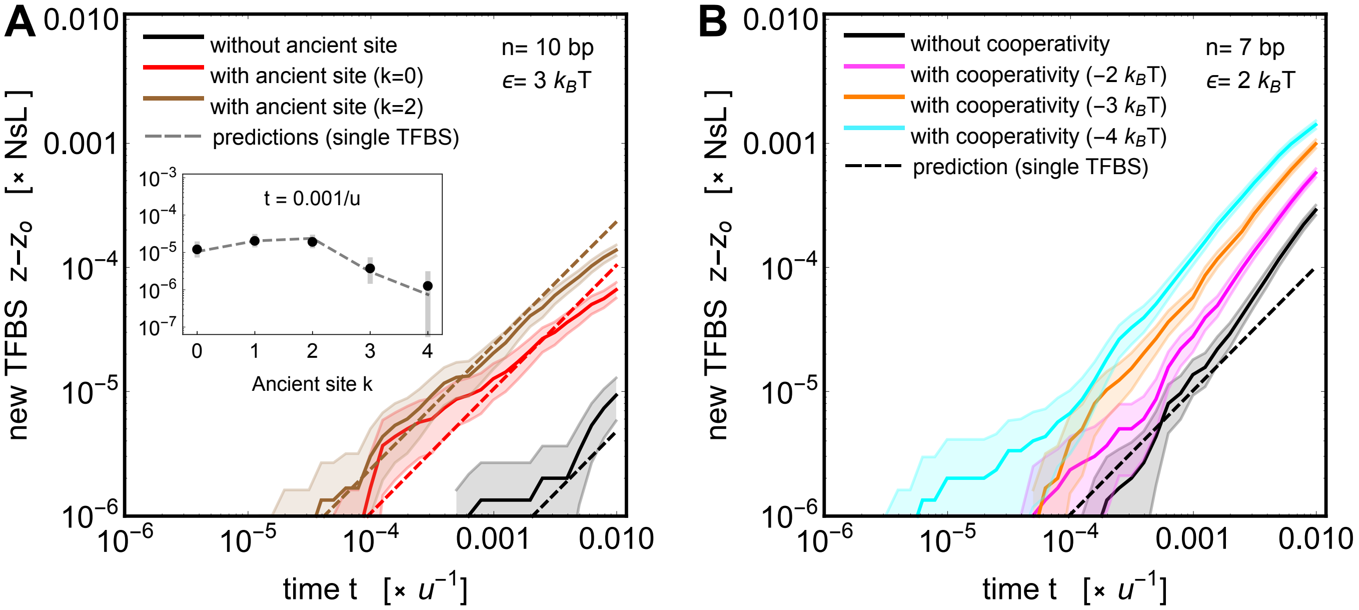

FIG. 5: **Ancient sites and cooperativity can accelerate the emergence of TF binding sites in longer regulatory sequences. A)** The expected number of newly evolved TFBS in the presence (red and brown) or absence (black) of an ancient site, for binding site length $n = 10$ bp, and specificity, $\epsilon = 3\,k_BT$. In this example, the ancient site was a consensus site ($k = 0$) or two mismatches away from it ($k = 2$) that evolved under neutrality for $t' = 0.1/u$ prior to starting this simulation. Dashed lines show the predictions of a simple analytical model, Eq (30). The inset shows how the number of newly evolved TFBS at $t = 0.001/u$ scales with the mismatch of the ancient site k (plot markers: simulation means; error bars: two standard errors of the mean; dashed curve: prediction). **B)** The expected number of newly evolved TFBS without (black) and with cooperative interactions (for different cooperativity strengths, magenta: $E_c = -2\,k_BT$, yellow: $E_c = -3\,k_BT$, cyan: $E_c = -4\,k_BT$, see Eq (11) in the Models & Methods and text) for binding site length $n = 7$ bp, and specificity, $\epsilon = 2\,k_BT$. Both panels use $\mu = 4$ k_BT, strong selection ($Ns = 100$) and a combination of point and indel mutations ($\theta = 0.15$), acting on a regulatory sequence of length $L = 30$ bp. Thick solid lines show an average over 1000 simulation replicates, shading denotes ±2 SEM.

cooperativity (see the Models & Methods, Eq (11) and Fig 1D). The genotype of a nearby site will then influence whether a given site acts as a strongly or weakly binding site. The presence of a cooperative site acts as a local shift in the chemical potential, which changes the weak/strong threshold, so that an individually weak site can become a strongly binding site. Simulations using cooperative binding presented in Fig 5B illustrate how cooperativity can increase the speed of evolution. This is specifically effective for short binding sites of intermediate or low specificity, where a cooperative energy contribution can strongly influence the number of sites in the strong binding class (see Fig 13).

DISCUSSION

In this study, we aimed at a better theoretical understanding of which biophysical and population genetic factors influence the fast evolution of TFBSs in gene regulatory DNA, making sequence specific TF binding a plausible mechanism for the evolution of gene regulation and for generating phenotypic diversity. Following Berg et al. (2004) [32], we combined a biophysical model for TF binding with a simple population genetic model for the rate of sequence evolution. The key assumptions are that binding probability is determined by a thermodynamic equilibrium; that fitness depends linearly on binding probability; and that populations are typically homogeneous in genotype, and so evolve by substitution of single point and short insertion/deletion (indel) mutations. Remarkably, the biophysical and the evolutionary models take the same mathematical form: in the biophysical model, binding probability depends on the binding energy, relative to thermal fluctuations, βE, whilst in the evolutionary model, the chance that a mutation fixes depends on its selective advantage, relative to random sampling drift, Ns.

For single TFBS evolution, we calculated the average transition time between genotypes, the inverse being a measure for the speed of the evolution. Our results indicate that TFBS evolution is typically slow unless selection is very strong. It is important to emphasize that gaining a TFBS by point mutations under neutral evolution is very unlikely, contrasting with the belief in the current literature (e.g., [5, 13]). This is mainly due to Stone & Wray's argument that functional sites could readily be found by a random walk [31]; however, their argument assumed that individuals follow independent random walks, which grossly overestimates the rate of evolution (see MacArthur & Brookfield [29]). Indeed, fast rates of gaining a single TFBS require not only strong selection but also initial sequences in the mutational neighborhood of the functional sites. Especially, "presites," i.e. sequences 1 bp away from threshold sequences, can be crucial since they can evolve to functional sites by single mutations. Indel mutations can increase the rate of gaining a single TFBS from distant sequences, since they connect the genotype space extensively, but their effect is limited under realistic indel mutation rates [72, 73]. Future studies should consider the updates in estimates of indel mutation rates, since they are currently not as precise as point mutation rates, although we do not expect big qualitative departures from our results.

Considering the evolution of a single TFBS from random sequence, we showed that biophysical parameters, binding length and specificity, are constrained for realistic evolutionary gain rates from the most redundant mismatch classes. The rates drop exponentially with binding length, making TF whose binding length exceeds $10 - 12$ bp difficult to evolve from random sites, at least under the point and indel mutation mechanisms considered here. As a consequence of the biophysical fitness landscape, binding specificity and length show an inverse relation for the same magnitude of the gain rate from the most redundant mismatch class. Such an inverse relation is observed in position weight matrices of TFs collected from different databases for both eukaryotic and prokaryotic organisms, by Stewart & Plotkin (2012) [8]. In the same study, they reproduce this observation using a simple model which assumes that a trade-off between the selective advantage of binding to target sites, versus the selective disadvantage of binding to non-target sequence. Their model assumes a stationary distribution, and that sites are functional if they are mismatched at no more than one base. It would be interesting to explore a broader range of models that account for the dynamical coevolution between transcription factor binding specificity, its length, and its binding sites [9]. One idea can be to combine the evolutionary dynamical constraints (against large binding length and high specificity, which we show here) with simple physical constraints of TF dilution in non-target DNA (against short binding length and low specificity, again in an inverse relation [44]).

For a single TF binding site, the stationary distribution for the mismatch with the consensus binding sequence depends on the binding energy, but also on the sequence entropy – that is, the number of sequences at different distances from the consensus. Typically, the distribution is bimodal: either the site is functional, and is maintained by selection, or it is non-functional, and evolves almost neutrally. We show that it may take an extremely long time for the stationary distribution to be reached. Functional sites are unlikely to be lost if selection is strong (i.e., $Ns \gg 1$), whilst function is unlikely to evolve from a random sequence by neutral evolution, even if predicted under stationarity assumption. Therefore, typical rapid convergence to stationary distribution should be considered with caution in theoretical studies.

We showed that the dynamics of TFBS evolution in longer DNA sequences can be understood from the dynamics of single TFBS. The rate of evolution of new binding sites will be accelerated in proportion to the length of the promoter/enhancer sequence in which that can be functional; however, because this increase is linear in promoter/enhancer length, it will have a weaker influence than the exponential effect changes in specificity or length of binding site. Especially the earlier dynamics (relevant for speciation timescales) are determined by the availability of presite biased sequences. Any process that allowed selection to pick up more distant sequences or that increased presite ratio among non-functional sites would accelerate adaptation from "virgin" sequences.

A key factor for an enrichment in presite ratio may arise through variation in GC content or through simple sequence repeats (especially if the preferred sequence has some repetitive or palindromic structure). In this study, we showed that it may also arise from ancient sites, i.e. sites that were functional in earlier evolutionary history and decayed into nonfunctional classes in evolution. Since loss of function is slow (comparable to the neutral mutation rate once selection becomes ineffective), this is plausible for sites that are under intermittent selection, or where there is a shift to binding by a new TF with similar preferred sequence [47, 85]. This effect of the earlier evolution can be especially important for long binding TFs as convergence to a truly randomized sequence distribution requires much longer times. MacArthur and Brookfield [29] showed that real promoter sequences may acquire functional sites more quickly than random sequence, but it is not clear whether that is due to a different general composition, or to the ghosts of previous selection. New studies are required to test our enriched presite-biased sequence hypothesis, especially for orthologous regions where functional TFBS is observed in sister populations or species. In a recent study, Villar *et al.* (2015) [54] provide evidence that enhancer DNA sequence structure is older than other DNA portions, suggesting the reuse of such regions in evolution, plausibly by gaining and losing TFBSs in repetitive manner. Nourmohammad & Lassig (2011) [30] showed evidence suggesting that local duplication of sequences followed by point mutations played important role in binding site evolution in Drosophila species (but surprisingly, not in yeast species). Another interesting option would be the existence of "mobile" presites or their fragments, e.g., as sequences embedded into transposable elements that could be inserted before the gene under selection for high expression [25]. Presites can be considered as concrete examples of cryptic sequences [86], potential source of future diversity and evolvability. We believe that understanding the effects of presites would contribute to the predictability of genetic adaptations regarding gene regulation, especially in important medical applications such as antibiotic resistance or virus evolution.

We also showed that the evolution of a functional binding site in longer DNA can be accelerated by cooperativity between adjacent transcription factors. When a TF occupies a co-binding site, sufficient transcriptional activity can be achieved from sequences of larger mismatch classes, an effect similar to a local increase in TF concentration. This mechanism permits faster evolution towards strongly binding sequences, and seems most effective for short TFBS where it creates a selection gradient already in the redundancy rich mismatch classes. Cooperative physical interactions might allow the evolution of binding occupancy and thus expression without large underlying sequence changes, which might be a reason for the observed weak correlation between sequence and binding evolution at certain regulatory regions. Importantly, TFBS clustering in eukaryotic enhancers can be a consequence of the fast evolution with cooperativity, as also supported by a recent empirical study [11].

Our theoretical framework is relevant more broadly for understanding the evolution of gene regulatory architecture. Since the speed of TFBS evolution from random sequences is proportional to NsL, our results suggest that population size N and the length of regulatory sequences L can compensate for each other in terms of the rate of adaptation. This is exactly what is observed: eukaryotes typically have longer regulatory DNA regions but small population sizes, while prokaryotes evolve TFBS within shorter regulatory sequence fragments but have large population sizes. Similarly, prokaryotes might have achieved longer TF binding lengths n, as large population size allowed them to overcome the exponential decrease in the gain rates with increasing n. If relevant, these observations would suggest that an important innovation in eukaryotic gene regulation must have been the ability of the transcriptional machinery to integrate the simultaneous occupancy of many low-specificity transcription factors bound over hundreds of basepairs of regulatory sequence, a process for which we currently have no good biophysical model.

ACKNOWLEDGMENTS

We thank Magdalena Steinrück, Georg Rieckh, Ferran Palero and Ziya Kalay for helpful comments. NB acknowledges support from ERC Advanced Grant 250152 "Selection and Information."

REFERENCES

[1] Fay JC, Wittkopp PJ. Evaluating the role of natural selection in the evolution of gene regulation. Heredity. 2007;100:191–199.

[2] Zheng W, Gianoulis TA, Karczewski KJ, Zhao H, Snyder M. Regulatory Variation Within and Between Species. Annual Review of Genomics and Human Genetics. 2011;12(1):327–346.

[3] Romero IG, Ruvinsky I, Gilad Y. Comparative studies of gene expression and the evolution of gene regulation. Nature Reviews Genetics. 2012 Jul;13(7):505–516.

[4] Hoekstra HE, Coyne JA. The locus of evolution: evo devo and the genetics of adaptation. Evolution; International Journal of Organic Evolution. 2007 May;61(5):995–1016.

[5] Wittkopp PJ. Evolution of Gene Expression. In: The Princeton Guide to Evolution. Princeton University Press; 2013. p. 413–419.

[6] Yao P, Lin P, Gokoolparsadh A, Assareh A, Thang MWC, Voineagu I. Coexpression networks identify brain region-specific enhancer RNAs in the human brain. Nature Neuroscience. 2015 Aug;18(8):1168–1174.

[7] Wunderlich Z, Mirny LA. Different gene regulation strategies revealed by analysis of binding motifs. Trends in genetics. 2009 Oct;25(10):434–440.

[8] Stewart AJ, Plotkin JB. Why transcription factor binding sites are ten nucleotides long. Genetics. 2012 Nov;192(3):973–985.

[9] Lynch M, Hagner K. Evolutionary meandering of intermolecular interactions along the drift barrier. Proceedings of the National Academy of Sciences of the United States of America. 2015. 112:E30-E38.

[10] Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, et al. Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. Science. 2010 May;328(5981):1036–1040.

[11] Stefflova K, Thybert D, Wilson M, Streeter I, Aleksic J, Karagianni P, et al. Cooperativity and Rapid Evolution of Cobound Transcription Factors in Closely Related Mammals. Cell. 2013 Aug;154(3):530–540.

[12] Dowell RD. Transcription factor binding variation in the evolution of gene regulation. Trends in Genetics. 2010 Nov;26(11):468–475.

[13] Villar D, Flicek P, Odom DT. Evolution of transcription factor binding in metazoans - mechanisms and functional implications. Nature Reviews Genetics. 2014 Apr;15(4):221–233.

[14] Doniger SW, Fay JC. Frequent Gain and Loss of Functional Transcription Factor Binding Sites. PLoS Comput Biol. 2007 May;3(5):e99.

[15] Moses AM, Pollard DA, Nix DA, Iyer VN, Li XY, Biggin MD, et al. Large-Scale Turnover of Functional Transcription Factor Binding Sites in Drosophila. PLoS Comput Biol. 2006 Oct;2(10):e130.

[16] Ludwig MZ, Patel NH, Kreitman M. Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. Development. 1998;p. 949–958.

[17] Paris M, Kaplan T, Li XY, Villalta JE, Lott SE, Eisen MB. Extensive Divergence of Transcription Factor Binding in Drosophila Embryos with Highly Conserved Gene Expression. PLoS Genet. 2013 Sep;9(9):e1003748.

[18] Ellison CE, Bachtrog D. Dosage Compensation via Transposable Element Mediated Rewiring of a Regulatory Network. Science. 2013 Nov;342(6160):846–850.

[19] Alekseyenko AA, Ellison CE, Gorchakov AA, Zhou Q, Kaiser VB, Toda N, et al. Conservation and de novo acquisition of dosage compensation on newly evolved sex chromosomes in Drosophila. Genes & Development. 2013 Apr;27(8):853–858.

[20] Contente A, Dittmer A, Koch MC, Roth J, Dobbelstein M. A polymorphic microsatellite that mediates induction of PIG3 by p53. Nature Genetics. 2002 Mar;30(3):315–320.

[21] Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, et al. Variation in Transcription Factor Binding Among Humans. Science. 2010 Apr;328(5975):232–235.

[22] Chan YF, Marks ME, Jones FC, Villarreal G, Shapiro MD, Brady SD, et al. Adaptive Evolution of Pelvic Reduction in Sticklebacks by Recurrent Deletion of a Pitx1 Enhancer. Science. 2010 Jan;327(5963):302–305.

[23] Vierstra J, Rynes E, Sandstrom R, Zhang M, Canfield T, Hansen RS, et al. Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. Science. 2014 Nov;346(6212):1007–1012.

[24] Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. Variable Tandem Repeats Accelerate Evolution of Coding and Regulatory Sequences. Annual Review of Genetics. 2010;44(1):445–477.

[25] Feschotte C. Transposable elements and the evolution of regulatory networks. Nature Reviews Genetics. 2008 May;9(5):397–405.

[26] Hahn MW, Stajich JE, Wray GA. The Effects of Selection Against Spurious Transcription Factor Binding Sites. Molecular Biology and Evolution. 2003 Jun;20(6):901–906.

[27] He BZ, Holloway AK, Maerkl SJ, Kreitman M. Does Positive Selection Drive Transcription Factor Binding Site Turnover? A Test with Drosophila Cis-Regulatory Modules. PLoS Genet. 2011 Apr;7(4):e1002053.

[28] Arnold CD, Gerlach D, Spies D, Matts JA, Sytnikova YA, Pagani M, et al. Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. Nature Genetics. 2014 Jul;46(7):685–692.

[29] MacArthur S, Brookfield JFY. Expected Rates and Modes of Evolution of Enhancer Sequences. Molecular Biology and Evolution. 2004 Jun;21(6):1064–1073.

[30] Nourmohammad A, Lässig M. Formation of Regulatory Modules by Local Sequence Duplication. PLoS Comput Biol. 2011 Oct;7(10):e1002167.

[31] Stone JR, Wray GA. Rapid evolution of cis-regulatory sequences via local point mutations. Molecular Biology and Evolution. 2001 Sep;18(9):1764–1770.

[32] Berg J, Willmann S, Lässig M. Adaptive evolution of transcription factor binding sites. BMC Evolutionary Biology. 2004 Oct;4(1):42.

[33] von Hippel PH, Berg OG. On the specificity of DNA-protein interactions. Proceedings of the National Academy of Sciences of the United States of America. 1986 Mar;83(6):1608–1612.

[34] Berg OG, von Hippel PH. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. Journal of molecular biology. 1987 Feb;193(4):723–750.

[35] Stormo GD, Fields DS. Specificity, free energy and information content in protein-DNA interactions. Trends in biochemical sciences. 1998 Mar;23(3):109–113.

[36] Stormo GD, Hartzell GW. Identifying protein-binding sites from unaligned DNA fragments. Proceedings of the National Academy of Sciences. 1989 Feb;86(4):1183–1187.

[37] Stormo GD, Zhao Y. Determining the specificity of protein-DNA interactions. Nature Reviews Genetics. 2010 Nov;11(11):751–760.

[38] Zhao Y, Granas D, Stormo GD. Inferring Binding Energies from Selected Binding Sites. PLoS Comput Biol. 2009 Dec;5(12):e1000590.

[39] Shea MA, Ackers GK. The OR Control system of bacteriophage lambda: A physical-chemical model for gene regulation. Journal of Molecular Biology. 1984;p. 211–230.

[40] Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, et al. Transcriptional regulation by the numbers: applications. Current Opinion in Genetics & Development. 2005;15:125–135.

[41] Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, et al. Transcriptional regulation by the numbers: models.

Current Opinion in Genetics & Development. 2005;15:116–124.

[42] Hermsen R, Tans S, ten Wolde PR. Transcriptional Regulation by Competing Transcription Factor Modules. PLoS Comput Biol. 2006 Dec;2(12):e164.

[43] Hermsen R, Ursem B, ten Wolde PR. Combinatorial Gene Regulation Using Auto-Regulation. PLoS Comput Biol. 2010 Jun;6(6):e1000813.

[44] Gerland U, Moroz JD, Hwa T. Physical constraints and functional characteristics of transcription factor-DNA interaction. Proceedings of the National Academy of Sciences of the United States of America. 2002 Sep;99(19):12015–12020.

[45] Gerland U, Hwa T. On the selection and evolution of regulatory DNA motifs. Journal of Molecular Evolution. 2002 Oct;55(4):386–400.

[46] Stewart AJ, Plotkin JB. The evolution of complex gene regulation by low-specificity binding sites. Proceedings of the Royal Society B: Biological Sciences. 2013 Oct;280(1768).

[47] Payne JL, Wagner A. The Robustness and Evolvability of Transcription Factor Binding Sites. Science. 2014 Feb;343(6173):875–877.

[48] Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U. Predicting expression patterns from regulatory sequence in Drosophila segmentation. Nature. 2008 Jan;451(7178):535–540.

[49] Samee MAH, Sinha S. Quantitative Modeling of a Gene's Expression from Its Intergenic Sequence. PLoS Comput Biol. 2014 Mar;10(3):e1003467.

[50] He X, Samee AH, Blatti C, Sinha S. Thermodynamics-Based Models of Transcriptional Regulation by Enhancers: The Roles of Synergistic Activation, Cooperative Binding and Short-Range Repression. PLOS Computational Biology. 2010.

[51] He X, Duque TSPC, Sinha S. Evolutionary Origins of Transcription Factor Binding Site Clusters. Molecular Biology and Evolution. 2012 Mar;29(3):1059–1070.

[52] Duque T, Samee MAH, Kazemian M, Pham HN, Brodsky MH, Sinha S. Simulations of Enhancer Evolution Provide Mechanistic Insights into Gene Regulation. Molecular Biology and Evolution. 2013 Oct;31(1):184–200.

[53] Duque T, Sinha S. What Does It Take to Evolve an Enhancer? A Simulation-Based Study of Factors Influencing the Emergence of Combinatorial Regulation. Genome Biology and Evolution. 2015 Jun;7(6):1415–1431.

[54] Villar D, Berthelot C, Aldridge S, Rayner T, Lukk M, Pignatelli M, et al. Enhancer Evolution across 20 Mammalian Species. Cell. 2015 Jan;160(3):554–566.

[55] Desai MM, Fisher DS. Beneficial Mutation-Selection Balance and the Effect of Linkage on Positive Selection. Genetics. 2007 Jul;176(3):1759–1798.

[56] Lynch M, Conery JS. The Origins of Genome Complexity. Science. 2003 Nov;302(5649):1401–1404.

[57] Kimura M. On the Probability of Fixation of Mutant Genes in a Population. Genetics. 1962 Jun;47(6):713–719.

[58] Hammar P, Wallden M, Fange D, Persson F, Baltekin Ö, Ullman G, et al. Direct measurement of transcription factor dissociation excludes a simple operator occupancy model for gene regulation. Nature Genetics. 2014 Apr;46(4):405–408.

[59] Cepeda-Humerez SA, Rieckh G, Tkačik G. Stochastic proofreading mechanism alleviates crosstalk in transcriptional regulation. arXiv:150405716 [q-bio]. 2015 Apr;ArXiv: 1504.05716. Available from: http://arxiv.org/abs/1504.05716.

[60] Brewster RC, Jones DL, Phillips R. Tuning Promoter Strength through RNA Polymerase Binding Site Design in Escherichia coli. PLoS Computational Biology. 2012 Dec;8(12).

[61] Razo-Mejia M, Boedicker JQ, Jones D, DeLuna A, Kinney JB, Phillips R. Comparison of the theoretical and real-world evolutionary potential of a genetic circuit. Physical Biology. 2014 Apr;11(2):026005.

[62] Haldane A, Manhart M, Morozov AV. Biophysical Fitness Landscapes for Transcription Factor Binding Sites. PLoS Comput Biol. 2014 Jul;10(7):e1003683.

[63] McKeown AN, Bridgham JT, Anderson DW, Murphy MN, Ortlund EA, Thornton JW. Evolution of DNA Specificity in a Transcription Factor Family Produced a New Gene Regulatory Module. Cell. 2014 Sep;159(1):58–68.

[64] Weinert FM, Brewster RC, Rydenfelt M, Phillips R, Kegel WK. Scaling of Gene Expression with Transcription-Factor Fugacity. Physical Review Letters. 2014 Dec;113(25):258101.

[65] Maerkl SJ, Quake SR. A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors. Science. 2007 Jan;315(5809):233–237.

[66] Kinney JB, Murugan A, Callan CG, Cox EC. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. Proceedings of the National Academy of Sciences. 2010 May;107(20):9158–9163.

[67] Fields DS, He Yy, Al-Uzri AY, Stormo GD. Quantitative specificity of the Mnt repressor 1. Journal of Molecular Biology. 1997 Aug;271(2):178–194.

[68] Mirny LA. Nucleosome-mediated cooperativity between transcription factors. Proceedings of the National Academy of Sciences. 2010 Dec;107(52):22534–22539.

[69] Taylor MS, Ponting CP, Copley RR. Occurrence and Consequences of Coding Sequence Insertions and Deletions in Mammalian Genomes. Genome Research. 2004 Apr;14(4):555–566.

[70] Brandström M, Ellegren H. The Genomic Landscape of Short Insertion and Deletion Polymorphisms in the Chicken (Gallus gallus) Genome: A High Frequency of Deletions in Tandem Duplicates. Genetics. 2007 Jul;176(3):1691–1701.

[71] Park L. Ancestral Alleles in the Human Genome Based on Population Sequencing Data. PLoS ONE. 2015 May;10(5):e0128186.

[72] Cartwright RA. Problems and Solutions for Estimating Indel Rates and Length Distributions. Molecular Biology and Evolution. 2009 Feb;26(2):473–480.

[73] Chen JQ, Wu Y, Yang H, Bergelson J, Kreitman M, Tian D. Variation in the Ratio of Nucleotide Substitution and Indel Rates across Genomes in Mammals and Bacteria. Molecular Biology and Evolution. 2009 Jul;26(7):1523–1531.

[74] Lee H, Popodi E, Tang H, Foster PL. Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. Proceedings of the National Academy of Sciences. 2012 Oct;109(41):E2774–E2783.

[75] Keightley PD, Johnson T. MCALIGN: Stochastic Alignment of Noncoding DNA Sequences Based on an Evolutionary Model of Sequence Evolution. Genome Research. 2004 Mar;14(3):442–450.

[76] Wright S. Evolution in Mendelian Populations. Genetics. 1931 Mar;16(2):97–159.

[77] Sella G, Hirsh AE. The application of statistical physics to evolutionary biology. Proceedings of the National Academy of Sciences of the United States of America. 2005;102:9541–9546.

[78] Mustonen V, Lässig M. Evolutionary population genetics of promoters: Predicting binding sites and functional phylogenies. Proceedings of the National Academy of Sciences of the United States of America. 2005 Nov;102(44):15936–15941.

[79] Mustonen V, Kinney J, Callan CG, Lässig M. Energy-dependent fitness: A quantitative model for the evolution of yeast transcription factor binding sites. Proceedings of the National Academy of Sciences of the United States of America. 2008 Aug;105(34):12376–12381.

[80] Barton NH, Coe JB. On the application of statistical physics to evolutionary biology. Journal of Theoretical Biology. 2009 Jul;259(2):317–324.

[81] Manhart M, Haldane A, Morozov AV. A universal scaling law determines time reversibility and steady state of substitutions under selection. Theoretical Population Biology. 2012 Aug;82(1):66–76.

[82] Paixão T, Heredia JP, Sudholt D, Trubenova B. First Steps Towards a Runtime Comparison of Natural and Artificial Evolution. In: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO 2015, Madrid, Spain, July 11-15, 2015. ACM; 2015. p. 1455–1462.

[83] Otto SP, Day T. A Biologist's Guide to Mathematical Modeling in Ecology and Evolution. Princeton University Press; 2007.

[84] Giorgetti L, Siggers T, Tiana G, Caprara G, Notarbartolo S, Corona T, et al. Noncooperative Interactions between Transcription Factors and Clustered DNA Binding Sites Enable Graded Transcriptional Responses to Environmental Inputs. Molecular Cell. 2010 Feb;37(3):418–428.

[85] Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014 Sep;158(6):1431–1443.

[86] Rajon E, Masel J. Compensatory Evolution and the Origins of Innovations. Genetics. 2013 Jan;193(4):1209–1220.

SUPPORTING INFORMATION

Other fitness models for comparison & for interacting TFBSs

Power-law decaying fitness models for comparison:

In order to understand the importance of the thermodynamically-motivated sigmoid shape for the binding probability, we compare our results to those obtained with power-law functions that decay with exponent γ (note that $\gamma = \infty$ corresponds to a step-like fitness landscape), formally defined as

$$\pi_{\mathrm{pl}}(k) = \begin{cases} \pi_{\mathrm{TD}}(k) & k \leq k_{\mathcal{S}} \\ \left(k_{\mathcal{S}}/k\right)^{\gamma} \, \pi_{\mathrm{TD}}(k_{\mathcal{S}}) & k > k_{\mathcal{S}} \end{cases} . \tag{31}$$

Fig 8 shows that the power-law exponent is a major determinant of the gain rates, suggesting that a biophysically realistic fitness landscape is crucial for the quantitative understanding of TFBS evolution.

Fitness models of interacting TFBSs in larger regulatory sequence:

In addition to physical cooperativity between nearby TFs on promoter/enhancers (see the Models & Methods, Fig 5 and Fig 13), here we also consider two other models. The first additional model assumes that the binding occupancy of the strongest binding site in the regulatory sequence is the proxy for the gene expression level and the fitness, i.e.

$$f(\boldsymbol{\sigma}) = s \, \mathrm{MAX}\{\pi^{(\mathrm{i})}(\boldsymbol{\sigma})\}. \tag{32}$$

Note that different TFBSs interact with each other to compete for the strongest binding within a promoter or an enhancer.

The second additional model addresses synergistic interaction between the two strongest-binding TFBS, located anywhere in the regulatory sequence. This example is a simplified version of a biophysical model where TFs, binding anywhere in a regulatory region, compete for the occupancy of that region with a nucleosome (for a more elaborative modeling framework, see Mirny (2010) [68]). We call this type of interaction between two TFs “non-physical” because TFs don’t interact directly; their interaction is effectively mediated by some other biophysical process. The probability of the joint occupancy of the two TFs at promoter or enhancer can be used as the proxy for gene expression level and the fitness, i.e.

$$f(\boldsymbol{\sigma}) = s \, \frac{e^{-\beta(\epsilon(k_1+k_2)-2\mu)}}{1 + e^{-\beta(\epsilon k_1-\mu)} + e^{-\beta(\epsilon k_2-\mu)} + e^{-\beta(\epsilon(k_1+k_2)-2\mu)}}, \tag{33}$$

where k_1 and k_2 correspond to the genotypes of two TFBSs with the smallest mismatches in the regulatory sequence.

Do these models yield different result for the emergence of strong binding sites from random sequences at early evolutionary times ($\sim$ speciation time scales), in comparison to our main model, where the sum of binding occupancies is used as a proxy for gene expression level [Eq(7) in the main text]? For typical biophysical parameters (binding lenght: $n = 7$ bp, binding specificity: $\epsilon = 2\ k_B T$ and chemical potential: $\mu = 4\ k_B T$), we show in Fig 14 that these modified models do not differ extensively from results of our main model.

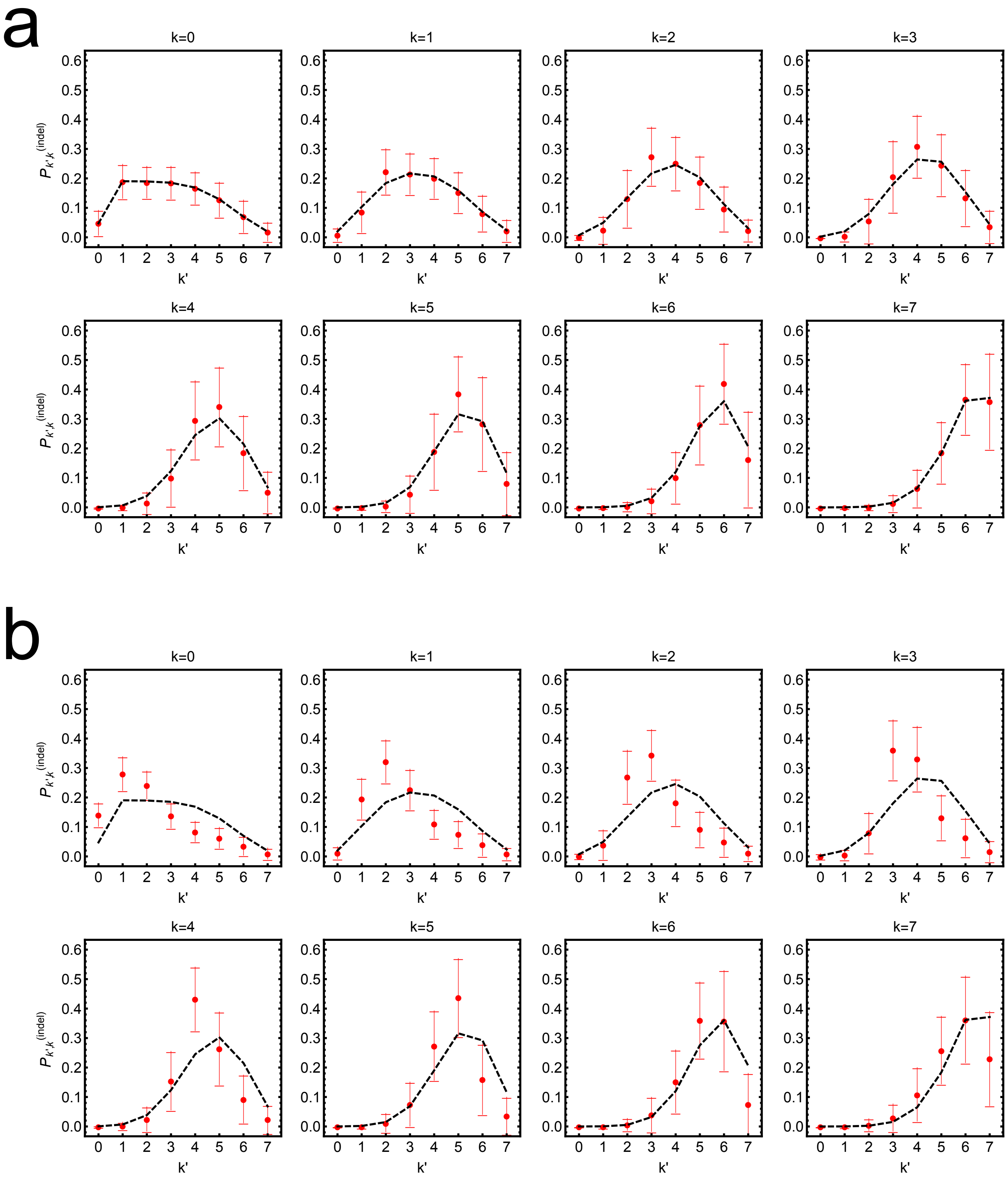

FIG. 6: **Indel mutations connect the mismatch genotype space differently from point mutations. a)** Probability that a binding site with k mismatches mutates to k' mismatches, for a single binding site of length $n = 7$ bp, according to our indel mutation model in a fixed genomic window (see the Models & Methods section). Dashed curve = analytical prediction according to Eq (13). Red points = mean ±1 std of 10^3 replicate realizations of the frequency distribution (for each replicate, 1 consensus sequence is created and 10^4 mutations are simulated for each k). **b)** The same analysis as in a), but allowing for a flexible genomic window for alignment after insertion mutations. We pick the minimal mismatch case to asses the quality of our approximation. As expected, this creates a bias towards smaller mismatch classes, but suggests that our approximation is still reasonable.

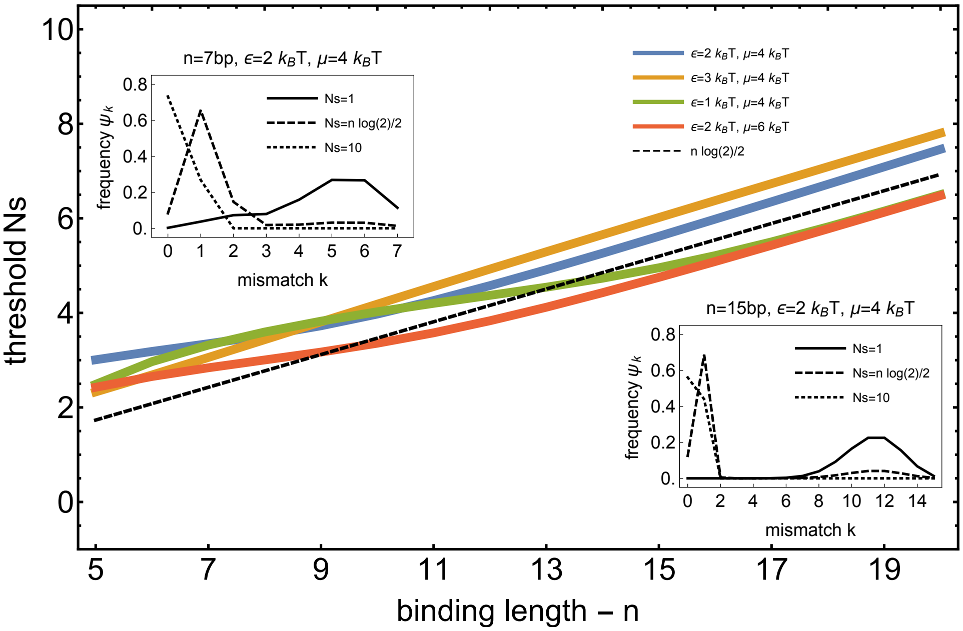

FIG. 7: **Threshold value of Ns for bimodality (i.e., threshold between strong and weak selection regimes).** The value of Ns at which 5% of the probability weight in the stationary distribution is in non-strong mismatch classes, i.e. $k > k_{\mathcal{S}}$. For selection stronger than this threshold, the stationary distribution is concentrated at low k (high fitness) classes and is practically unimodal. Different colors correspond to different biophysical parameters (see legend), analytical prediction $n\log(2)/2$ is in black (see the Models & Methods section and Eq (20)). Insets show examples of stationary distributions for different Ns values for short and long binding sites.

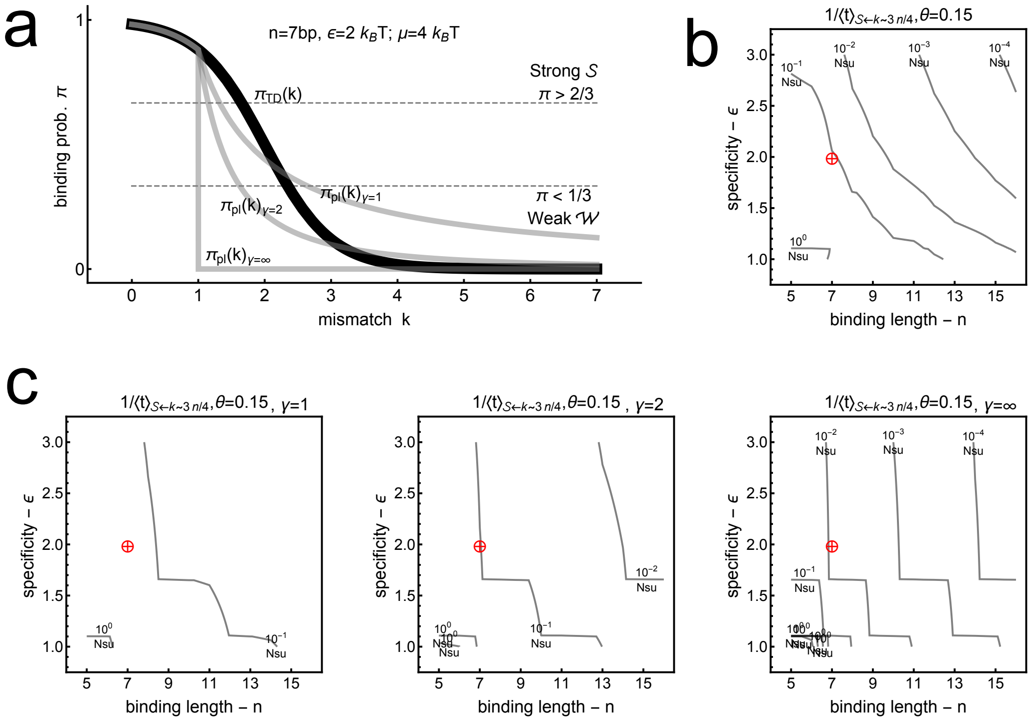

FIG. 8: **Single TFBS gain rates in modified fitness landscapes with a power-law tail.** The thermodynamic fitness landscape has been modified to have a power-law decaying tail of exponent γ for $k > k_{\mathcal{S}}$, as in Eq (31) in SI text. We tested $\gamma = 1$, 2 and ∞ corresponding to smooth, intermediate and step-like decay. Plot conventions are the same as in Fig 2C. **b)** Isolated TFBS gain rate from the most redundant mismatch class for the thermodynamic model, replotted from Fig 2C for reference. **c)** Plots analogous to b) using modified fitness landscapes defined by the power-law exponent γ. Gain rates are higher for small $\gamma = 1$ and lower for the step landscape ($\gamma = \infty$), relative to the reference.

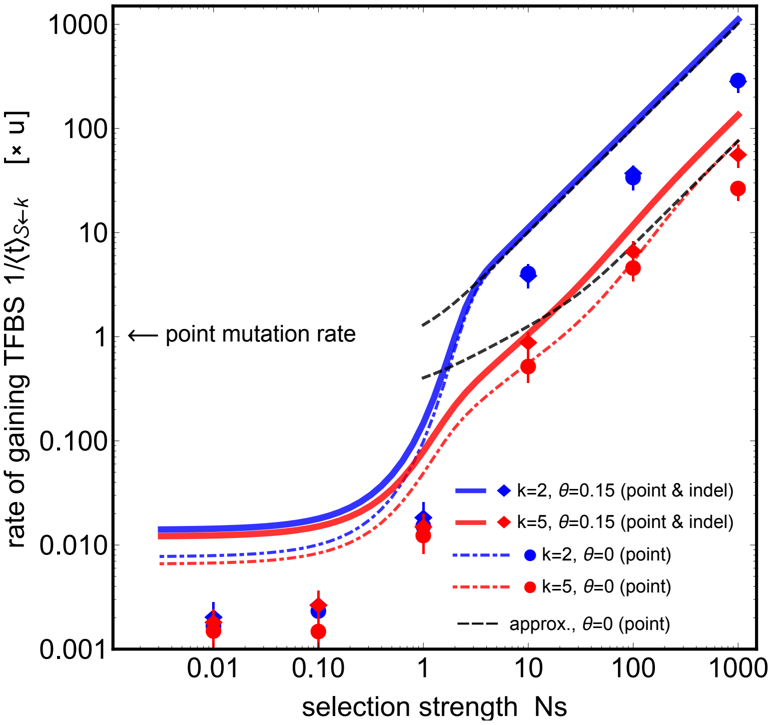

FIG. 9: **The effect of polymorphisms on the single TFBS gain rate at higher mutation rates.** Wright-Fisher simulation results (point markers, error bars = 2 standard errors of the mean) at $4Nu = 0.1$, in comparison to the fixed state model (continuous curves). Plot conventions are the same as in Fig 2. Biophysical parameters used: $n = 7$, $\epsilon = 2\ k_BT$, $\mu = 4\ k_BT$. Polymorphisms generally decrease TFBS gain rates.

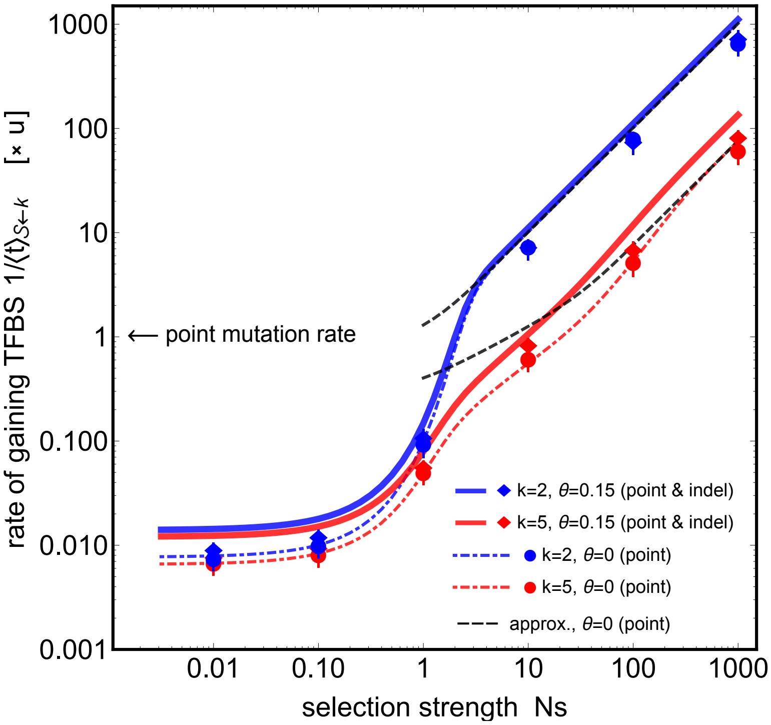

FIG. 10: **Relaxing the mismatch assumption.** Fig 2, but using energy matrices whose nonzero entries are gaussian random variables ε_i, such that $\langle \varepsilon_i \rangle = \epsilon = 2k_BT$ and $\sigma_\varepsilon = 0.5k_BT$; $n = 7$, $\mu = 4k_BT$. The analytical results under the equal mismatch assumption are shown in continuous lines.

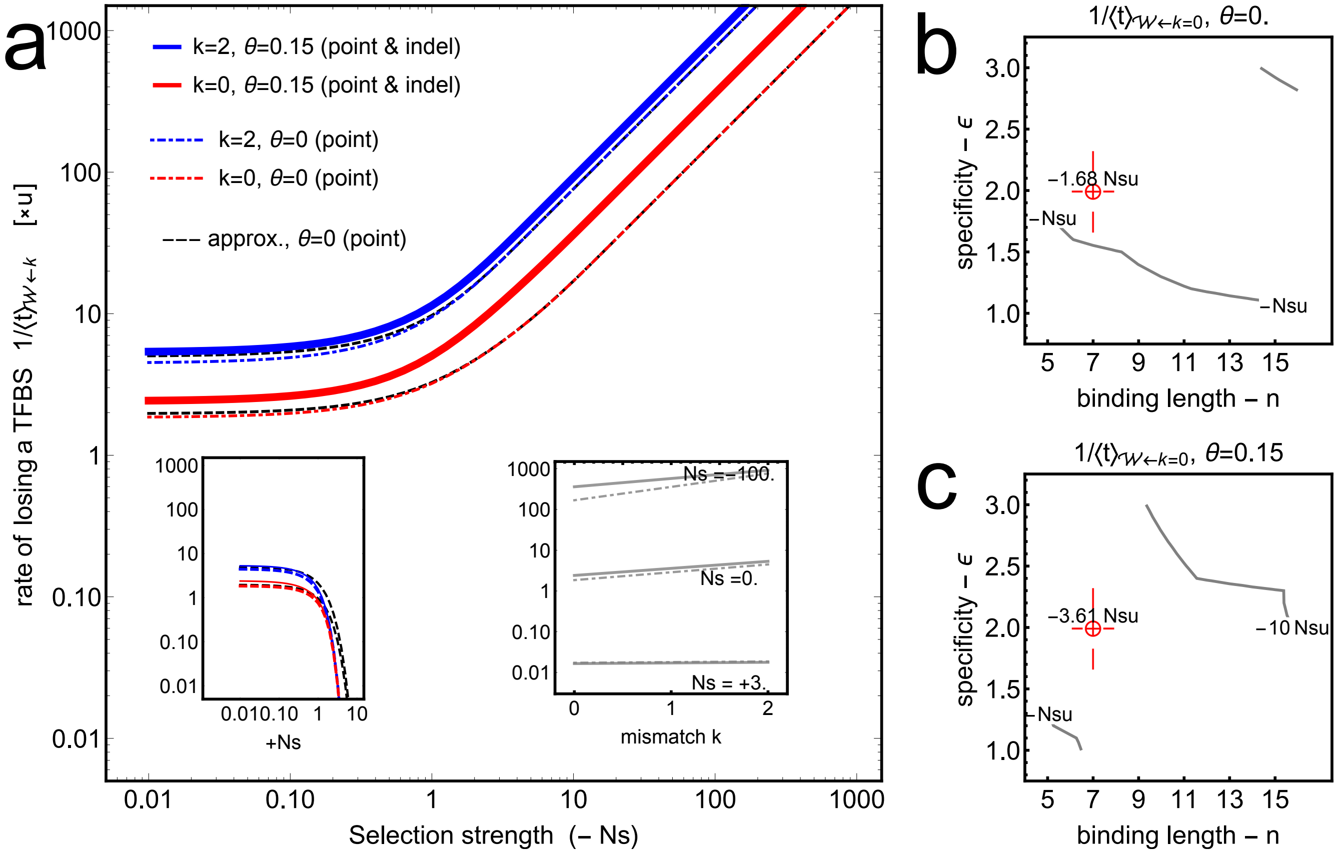

FIG. 11: **Single TF binding site loss rates at an isolated DNA region.** The dependence of the loss rate, $1/\langle t\rangle_{\mathcal{W}\leftarrow k}$ shown in units of point mutation rate, from sequences in different initial mismatch classes k (blue: $k = 2$, red: $k = 0$), as a function of negative selection strength. Results with point mutations only ($\theta = 0$) are shown by dashed line; with admixture of indel mutations ($\theta = 0.15$) by a solid line. For strong selection, $|Ns| \gg 1$, the rates scale with $2|Ns|nu$, which is captured well by the "shortest path" approximation (black dashed lines in the main figure) of Eq (24). The biophysical parameters are: site length $n = 7$ bp; binding specificity $\epsilon = 2\ k_BT$; chemical potential $\mu = 4\ k_BT$. Left inset: Ns-scaling with positive selection. Right inset: gain rates as a function of the initial mismatch class k for different Ns. **b, c)** Loss rates from the consensus sequence ($k = 0$) under strong negative selection, without (b) and with (c) indel mutations supplementing point mutations. Red crosshairs denote the cases depicted in panel a). Contour lines show constant loss rates in units of $Ns\,u$ as a function of biophysical parameters n and ϵ.

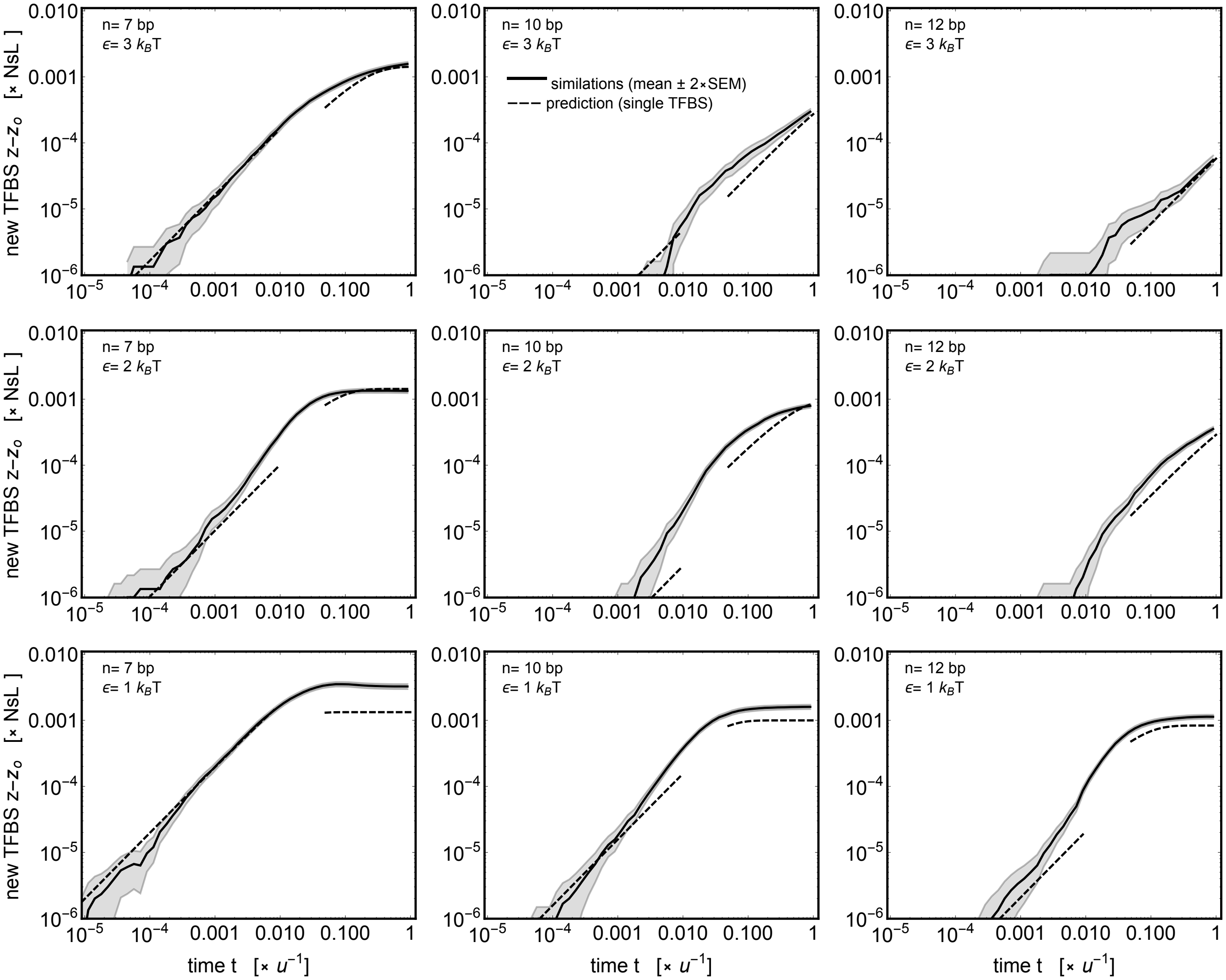

FIG. 12: **TFBS evolution in longer sequences**. Example simulations (black solid line) and analytic predictions based on single TFBS gain/loss rates (black dashed line), for different binding length n and specificity ϵ. Details are identical to Fig. 4.

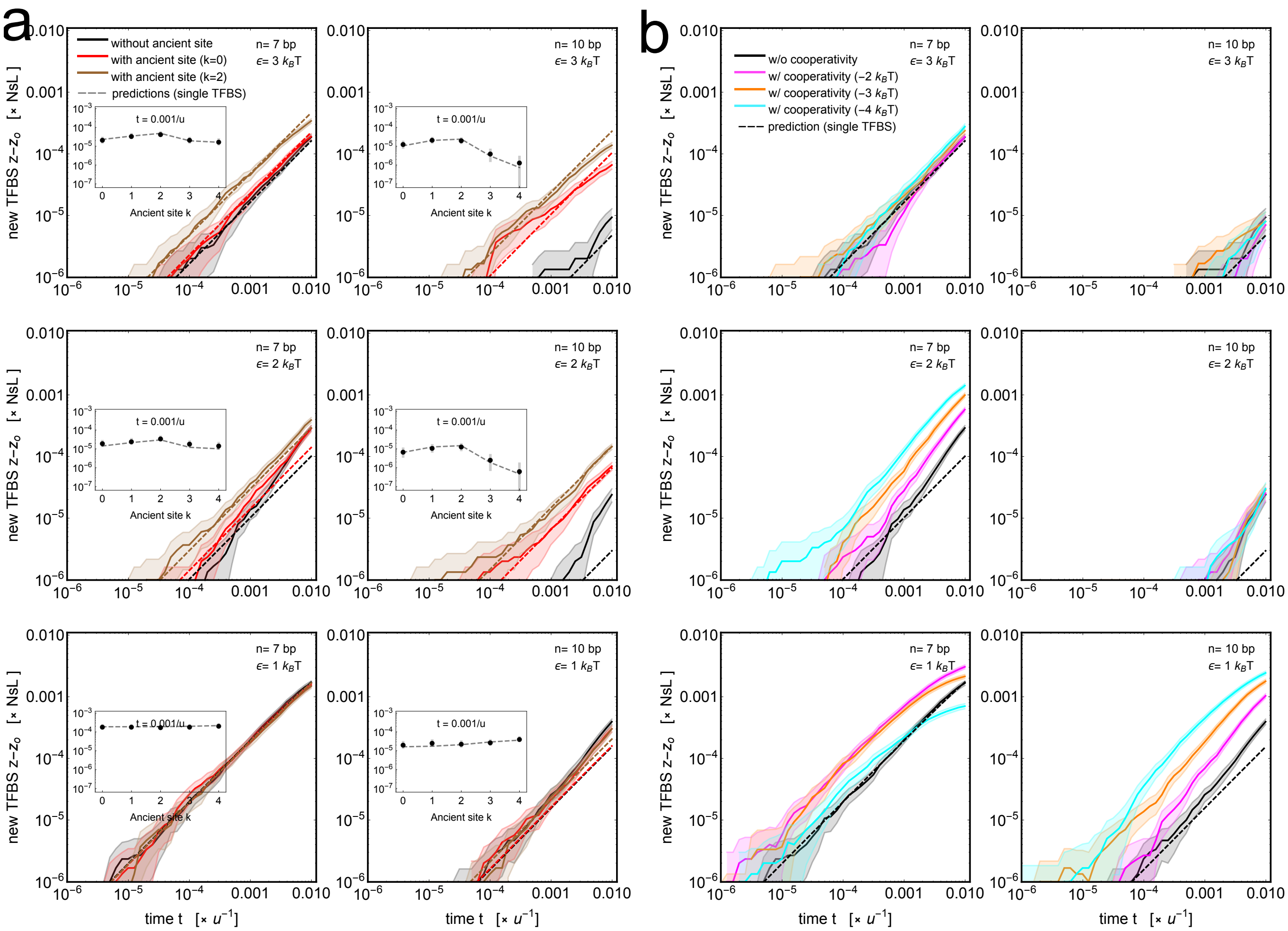

FIG. 13: **The effect of ancient sites (a) and cooperativity (b) for different binding lengths and specificities**. Simulations of TFBS evolution in longer sequences (colored lines) and analytic predictions based on single TFBS gain and loss rates (dashed black lines), analogous to Fig. 5. Different panels show different choices of TFBS binding length n and specificity ϵ. Ancient sites specifically facilitate the emergence of longer sites of high specificity, whereas cooperativity specifically facilitates the emergence of shorter sites of intermediate or low specificity.

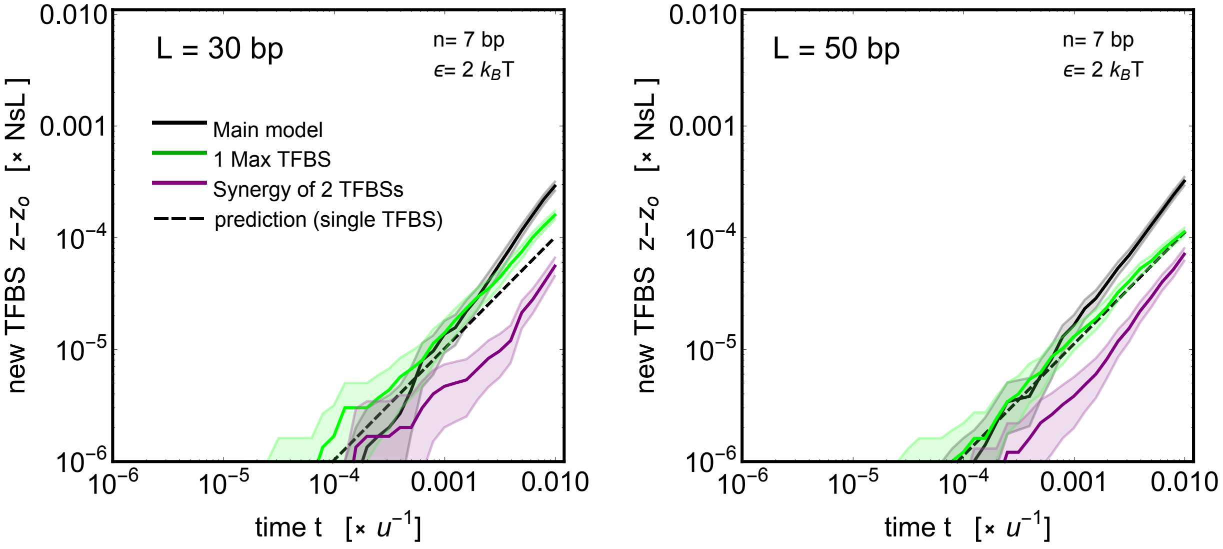

FIG. 14: **Fitness models of interacting TFBSs.** The expected number of newly evolved TFBS for binding site length $n = 7$ bp, specificity $\epsilon = 2\,k_BT$, and chemical potential $\mu = 4\,k_BT$ are shown for different fitness models. The solid black curve is the non-interacting model used in the main text (dashed curve: theoretical prediction). The green curve stands for the model of Eq (32) in SI text, where only the strongest binding site in the regulatory sequence determines gene expression. The purple curve stands for the model of Eq (33) in SI text, where two strongest TFBS synergistically determine the gene expression level. Shading denotes ±2 SEM. The simulations use regulatory sequences of length $L = 30$ bp (left) and $L = 50$ bp (right).

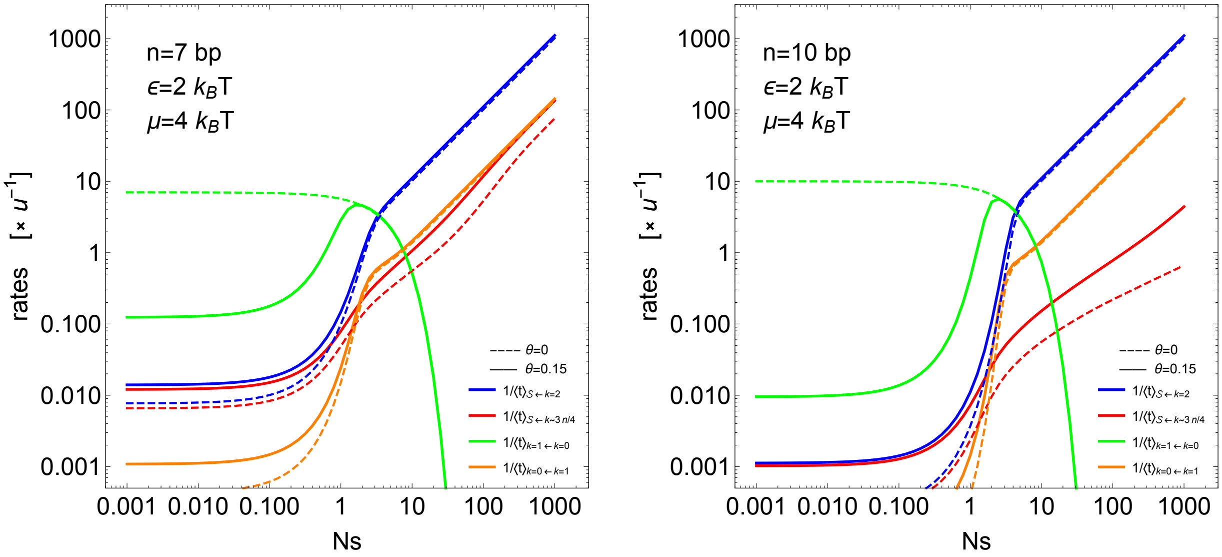

FIG. 15: **Comparison rates of TFBS gain rates and sequence turnover rates within functional TFBSs** Average first hitting times to particular mismatch k_j state can be calculated with a minor modification to Eq (21) by replacing $\mathcal{S}$ with k_j. The figures compare the rates of evolution of TFBS within the functional sites (i.e. $1/\langle t\rangle_{k=0\leftarrow k=1}$ and $1/\langle t\rangle_{k=1\leftarrow k=0}$). Plot conventions are the same as in Fig 2-A. Biophysical parameters used: $n = 7$ bp (left), $n = 10$ bp (right) $\epsilon = 2\ k_BT$, $\mu = 4\ k_BT$. It shows that for weak selection, the rates to evolve from $k = 0$ to $k = 1$ can be relatively faster. Also, although adaptation from random sites slows down with increasing n, we see that the adaptation rate to evolve from $k = 1$ to $k = 0$ can stay high.